# Evaluation of a prior-incorporated statistical model and established classifiers for externally visible characteristics prediction

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Maria-Alexandra Katsara

aus Athen, Griechenland

Köln, 2021

Datum der mündlichen Prüfung: 3 0 / 0 4 / 2 0 2 1

| | |
|---|---|
| Betreuer: | Herr Prof. Dr. Michael Nothnagel |
| Referent/in: | Herr Prof. Dr. Thomas Wiehe<br>Frau Prof. Dr. Juliette de Meaux |

# Table of Contents

# 1 Abbreviations

| | |
|---|---|
| STR | **Short tandem repeats** |
| FDP | **Forensic DNA phenotyping** |
| EVCs | **Externally visible characteristics** |
| ML | **Machine learning** |
| MLR | **Multinomial logistic regression** |
| SVM | **Support vector machines** |
| RF | **Random forest** |
| ANN | **Artificial neural networks** |
| AUC | **Area under curve** |
| PCR | **Polymerase chain reaction** |
| SNPs | **Single nucleotide polymorphisms** |
| NGS | **Next generation sequencing** |
| GWAS | **Genome wide association studies** |
| CODIS | **Combined DNA index system** |
| BGA | **Biogeographical ancestry** |
| AIC | **Akaike information criterion** |
| MPB | **Male pattern baldness** |
| GIANT | **Genetic investigation of anthropometric traits** |
| RBF | **Radial basis function** |
| PPV | **Positive predictive value** |
| NPV | **Negative predictive value** |

# 2 List of figures



# 3 List of tables

# 4 List of Publications

The present PhD thesis is based on the following list of publications. The discussion refers to the main publications and the main investigations that are not yet published.

## 1.1 Main publications

**Katsara M.A.**, Nothnagel M. 2019. **True colors: A literature review on the spatial distribution of eye and hair pigmentation**, Forensic Science International Genetics, 39:109-118, published online January 02, 2019. Doi: https://doi.org/10.1016/j.fsigen.2019.01.001

**Katsara M.A.**, Branicki W., Pospiech E., Hysi P., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. **Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits**, Forensic Science International Genetics, 50:102412, published online November 03, 2020. Doi: https://doi.org/10.1016/j.fsigen.2020.102412

**Katsara M.A.**, Branicki W., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. **Evaluation of supervised machine-learning methods for predicting appearance traits from DNA**, submitted to Forensic Science International Genetics, under revision.

# 5 Meeting abstracts (Talks/Poster presentations)

**Katsara M.A.**, Nothnagel M. 2019. True colors: A literature review on the spatial distribution of eye and hair pigmentation, 47th European Mathematical Genetics Meeting, April 8-9, 2019, Dublin, Ireland.

**Katsara M.A.**, Nothnagel M. 2019. True colors: A literature review on the spatial distribution of eye and hair pigmentation, 15th. Jahrestagung der Deutschen Gesellschaft für Abstammungsbegutachtung (DGAB), June 27-29, 2019, Cologne, Germany.

**Katsara M.A.**, Branicki W., Pospiech E., Hysi P., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits, International Science of Forensic Genetics conference, September 9-13, 2019, Prague, Czech Republic.

**Katsara M.A.**, Branicki W., Pospiech E., Hysi P., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. Testing the impact of trait prevalence priors in Bayesian-based genetic prediction

modeling of human appearance traits, 48th European Mathematical Genetics Meeting (EMGM), April 16-17, 2020, Lausanne, Switzerland.

# 6 Awards

Award for the best talk presentation: Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits, 48th European Mathematical Genetics Meeting (EMGM), April 16-17, 2020, Lausanne, Switzerland.

# 7 Summary

Human identification through DNA has played an important role in forensic science and in the criminal justice system for decades. It is referring to the association of genetic data with a particular human being and has facilitated police investigations in cases such as the identification of suspected perpetrators from biological traces found at crime scenes, missing persons, or victims of mass disasters [1]. Currently there are two main methods developed: the genotyping through short tandem repeats (STR profiling) and the forensic DNA phenotyping (FDP). Despite the fact that these two methods are aiming in identifying a person through its genetic material, their approach and consequences that come up are completely different. STR profiling compares allele repeats at specific loci in DNA and aims at a match with already known to the police authorities DNA profiles, while FDP, which is the focus on the current study, aims in the prediction of appearance traits of an individual [2, 3]. In contrast with STR profiling, information that arise out of FDP cannot be used as sole evidence in the court [4].

The ability of predicting EVCs from DNA can be used as 'biological witnesses' that can only provide leads for the investigative authorities and subsequently narrow down a possible large set of potential suspects. The use of FDP begins a new era of 'DNA intelligence' and holds great promise especially in cases where individuals cannot be identified with the conventional method of STR profiling and also in cases where there is no additional knowledge on the sample donor. So far in FDP, traits such as eye, hair and skin color can be predicted reliably with high prediction accuracy and predictive models have already been forensically validated [5-7]. Regarding other appearance traits, the current lack of knowledge on the genetic markers responsible for their phenotypic variation and the lower predictability, especially of intermediate categories, has prevented FDP from being routinely implemented in the field of forensic science.

The majority of the predictive models developed for appearance trait prediction were based on multinomial logistic regression (MLR) while only few used other methods such as decision trees and neural networks. Machine learning (ML) approaches have become a widely used tool for classification problems in several fields and they are known for their potential to boost model performance and their ability to handle different and complex types of data [8]. However, within the context of predicting EVCs, a systematic and comparative analysis among different ML approaches that could possibly indicate methods that outperform the standard MLR, has not been conducted so far. In addition, incorporation of priors in the EVC prediction models that may have potential to improve the already existing approaches, has not been investigated in the context of forensics yet. These priors indicate the trait category prevalence values among biogeographic ancestry groups, and their use would allow us to leverage Bayesian statistics in order to build more powerful prediction models. In our case, incorporation of such priors in the model could reflect the additional information from all yet unknown causal genetic factors and act as proxies in the prediction model. Therefore, those two approaches were conducted throughout my PhD project in order to improve the already existing approaches of FDP which was the main aim of my study.

In the first study, I aimed to collect a comprehensive data set from previously published sources on the spatial distribution of different appearance traits. I conducted a literature review in order to assemble this information, which later on could be incorporated as priors in the EVCs prediction models. Due to the lack of available and reliable sources, our resulting data set contained only eye and hair color for mostly European countries. More specifically, I collected data on eye color from 16 European and Central Asian countries, while for hair color I collected data from seven European countries. For countries outside of Europe, where the variation is low, it was not possible to assemble trustworthy and population-representative data. Afterwards, I calculated the association

of those two traits and obtained a moderate association between them. Interpolation techniques were applied in order to infer trait prevalence values in at least neighboring countries. Resulting prevalences and interpolated values were presented in spatial maps.

The subject of the second study was to incorporate the trait prevalence values as priors in the prediction model. However, due to the lack of reliable data that was observed in the first study, the incorporation of the actual priors that would give us the actual insight of their impact in the EVC prediction was not feasible with the current existing knowledge and the available data. Therefore, I assessed the impact of priors across a grid that contained all possible values that priors can take, for a set of appearance traits including eye, hair, skin color, hair structure, and freckles. In this way, I aimed to assess potential pitfalls caused by misspecification of priors. Results were compared and evaluated with the corresponding prior-free' previously established prediction models. The effect of priors was demonstrated in the standard performance measurements, including area under curve (AUC) and overall accuracy. I found out that from all possible prior values, there is a proportion that shows potential in improving the prediction accuracy. However, possible misspecification of priors can significantly diminish the overall accuracy. Based on that, I emphasize the importance of accurate prior values in the prediction modelling in order to identify the actual impact. As a consequence of the above, the use of prior informed models in forensics is currently infeasible and more studies on the topic are necessary in order to extend the current knowledge on spatial trait prevalence.

Finally, the focus of the third study was exploring and comparing the performances of methodologies beyond MLR. MLR is considered the standard method for predicting EVCs, since the majority of the predictive models developed are based on that method. Due to the fact that there is still potential for improvement of MLR models, especially for traits such as skin color or hair structure, I aimed at applying different ML methods in order to identify whether there is a potential classifier that outperforms the conventional method of MLR. Therefore I conducted a systematic comparison between MLR and three alternative ML classifiers, namely support vector machines (SVM), random forests (RF) and artificial neural networks (ANN). The traits that I focused on here were eye, hair, and skin color. All models were based on the genetic markers that were previously established in IrisPlex, HIrisPlex and HIrisPlex-S [5-7]. Overall, I observed that all four classifiers performed almost equally well, especially for eye color. Only non-substantial differences were obtained across the different traits and across trait categories. Given this outcome, none of the ML methods applied here performed better than MLR, at least for the three traits of eye, hair, and skin color. Ultimately, due to the easier interpretability of the MLR, it is suggested at least for now and for the currently known marker sets, that the use of MLR is the most appropriate method for predicting appearance traits from DNA.

Throughout my PhD project, it became apparent that the available knowledge on spatial trait prevalence values was quite restricted not only in certain appearance traits but also in continental groups. More specifically, most available and reliable data were focused on European populations and the traits that were available were mostly for eye and hair color. For other traits, such as skin color, hair structure, and freckles, the data were either extremely few or nonexistent. This was a significant obstacle throughout the project, since it prevented me from applying and testing the actual impact of the accurate trait prevalence values as priors in EVC prediction. However, the lack of data presented an opportunity to perform in-depth theoretical research, in particular testing the impact of priors within a spatial grid that included its possible values. I found out that there is a proportion of priors that showed potential to improve EVC prediction. However, caution is advised regarding misspecification of priors that can significantly deteriorate the models' performance. Furthermore, the application of different ML

approaches did not show any significant improvement on the prediction performance against the standard MLR. This could be due to the nature of the traits, since some of them are multifactorial and affected by various external independent factors or due to possible limitations of the currently known predictive markers. With the available knowledge so far, it is emphasized throughout this study that for the time being, priors are refrained from being incorporated in the EVC prediction models while from the different classifiers applied, MLR is considered as the most appropriate method for EVC prediction due to its easier interpretability. In addition, the presented study highlights the importance of reference data on externally visible traits and the identification of more genetic markers that contribute to certain traits and I hope that the present work will motivate the emergence of these certain types of data collections that potentially may improve the current EVC prediction models.

# 8 Introduction

## 8.1 Human identification from DNA markers

Human identification through DNA markers is considered to be the gold standard in the field of forensic science. It is referring to the analysis of the genetic data of an individual, which creates a unique DNA profile and can provide likelihoods of the involvement of the individual in unsolved cases [1, 9]. Substantial and continuous effort has been made to identify missing persons, human remains after wars, and analyzing the remains found at crime scenes in order to identify potential trace donors. The biological samples found at a crime scene can lead the investigating authorities in obtaining a DNA profile that can be compared with potentially large groups of suspects and subsequently can narrow it down or even identify the trace donor.

Identification can be done by extracting the DNA evidence found mostly at blood stains, buccal swabs or body fluids [10]. Its history starts back in 1986 by Alec Jeffreys, who used it for the case of two rapes and murders that happened in 1983 and 1986 [11]. In these cases, which took place in the United Kingdom, the fingerprints were collected and connected with semen stains that were found at the crime scenes. In 1987, genetic fingerprinting was first used in the USA in the case of Tommie Lee Andrews, a rapist from Florida, who was caught and sentenced into 22 years in prison [12]. Another important case was the one of missing children in Argentina where Mary-Claire King compared genetic material from children who were kidnapped by the Argentinian military between 1970's and early 1980's. Ultimately, she identified 59 children and helped them return back to their biological families [13].

Thus far there are two main approaches developed, namely the STR profiling and the Forensic DNA Phenotyping (FDP). The first method aims at human identification by matching profiles in DNA databases. In other words it focuses on the quantification of how likely is to obtain such a match by chance [10]. On the other hand, FDP aims in the prediction of appearance traits that subsequently can aid police investigations by narrowing down a possible large set of suspects [3]. Results from STR profiling are used in court almost routinely and therefore exonerate people who were wrongly convicted and establish or exclude paternity, while FDP outcomes have not reached thus far and likely they will not reach a scientific consensus that will allow them to be used as sole evidence in criminal courts [4]. This is due to the several issues that surround FDP such as the existing uncertainties in predictions, the current insufficient spatial resolution for inferring biogeographic ancestry and the fact that some appearance traits can be easily covered or changed, such as eye and hair color. For the sake of completeness it is important to mention here, despite the two aforementioned methods, other markers have also been used. One example is the recently developed method of long-familial DNA searches which shows potential for DNA matching techniques. However, this is out of the scope of the current study and for further details the study of Erlich et al. is advised [14].

### 8.1.1 STR profiling for human identification

Genetic fingerprinting which was the earliest established method for routine DNA identification and is considered to be a gold standard, is based on genetic profiles from DNA polymorphisms (short tandem repeats [STRs] or microsatellites) [15]. STRs are DNA segments ranging from two to six bases that are repeated numerous times and are distributed abundantly through the DNA sequence [16]. In this method DNA is amplified via Polymerase Chain Reaction (PCR) and electrophoresis in order to target the sequence specific primers and subsequently construct a

DNA profile (**Figure 1**). The generated STR profile is then compared with database of DNA profiles (e.g. the U.S. national DNA database). The purpose of this method is either to exclude or confirm the identity of a suspected perpetrator and it can be used as evidence in the court. One example of a standard marker set that is currently used by FBI for STR profiling is the Combined DNA Index system (CODIS). CODIS is a software project that began in 1990 in FBI's laboratories and in 1994 was established for law enforcement purposes. It offers investigative leads in cases where biological traces are found and recovered in crime scenes and contains multiple databases regarding the type of information that is being searched. Such examples are information on missing persons, convicted offenders and forensic samples collected from crime scenes [10]. A set of 13 loci was initially included in CODIS software including CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX and vWA, while in January 2017 seven more loci were included, namely: D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 and D22S1045 [17, 18]. Another example of a marker set used for STR profiling is the AmpFISTR Profiler Plus kit (Applied Biosystems, Foster City, CA) which includes nine markers already included in CODIS (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820) and a gender identification locus, the Amelogenin, which has been already validated by FBI and SWGDAM guidelines [19]. Furthermore, the PowerPlex 16 BIO STR system which contains the 13 loci of CODIS, the Amelogenin and two pentanucleotide loci (Penta D and Penta E) and its extension to PowerPlex 21 BIO STR system that includes all CODIS loci and additional loci that are commonly used in Asia and Europe [20, 21].

The advantage of STR profiling is that these markers are highly informative due to their high allele diversity, making the probability of a random match 1 ~ 3 trillion [22, 23]. In practical terms, these values ensure that each individual, except from identical twins, has a unique genetic profile. However, the disadvantage of this method lies in the fact that it is a comparative method which always requires a DNA database that contains the suspect in order to obtain a match. In other words, it requires that the suspect has already been 'seen' by the investigative authorities. Furthermore, the base pairs of the repetitions are not always available since the DNA from crime scenes is most of the times degraded.

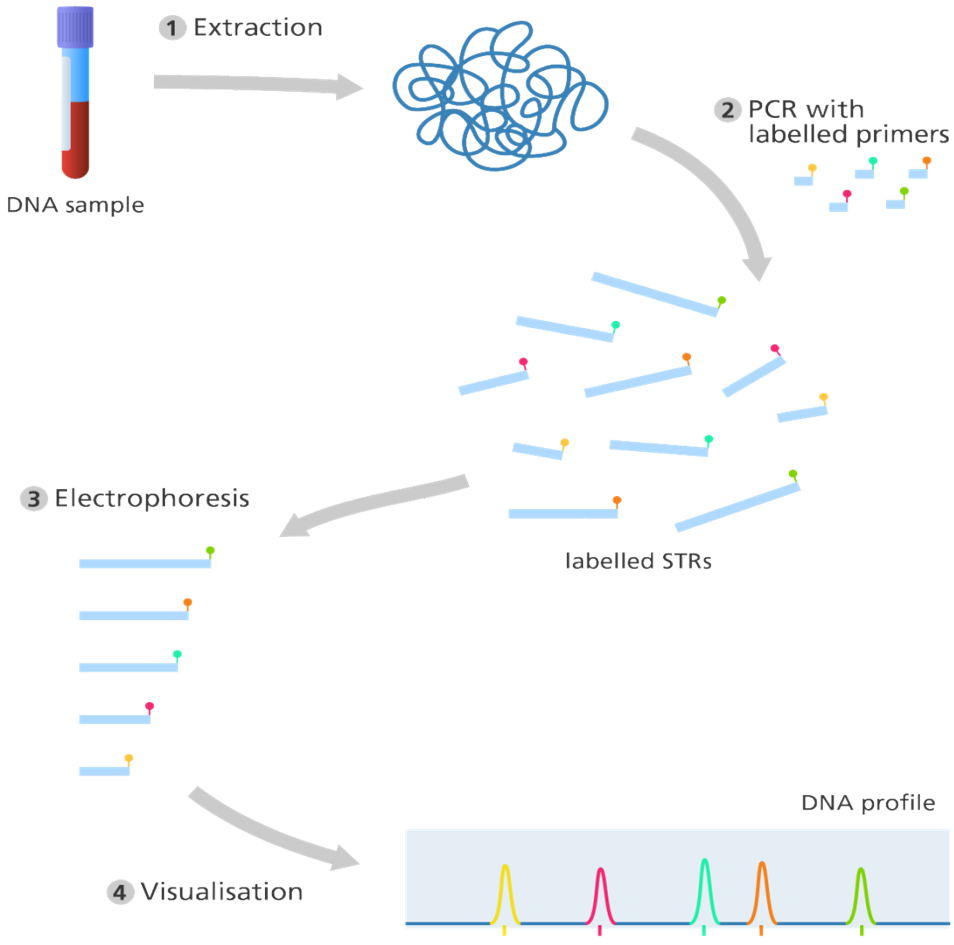


**Figure 1: (adapted from: yourgenome, (2016). Copyright information. [Online] Available at: https://www.yourgenome.org/copyright [Accessed 22.12.2020]):** Illustration showing the steps in DNA profiling. Image credit: Genome Research Limited

### 8.1.2 Forensic DNA Phenotyping for human identification

Forensic DNA Phenotyping (FDP), which is the main focus of the current study, is a method which is mostly based on the single nucleotide polymorphisms (SNPs) at least when it comes to the inference of appearance traits. A SNP is a variation in the DNA sequence which occurs when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome differs among individuals. If more than 1% of the population does not carry the same nucleotide at certain DNA position then this variation can be considered as SNP [24]. SNPs can influence visible characteristics and other phenotypes, since different alleles can result in different gene products or differential gene expression. They can be obtained from biological traces collected at crime scenes by either microarray genotyping or Next Generation Sequencing (NGS). Afterwards, selected genetic markers responsible for specific appearance traits that have been identified either via genome wide association studies (GWAS) or linkage analysis, are used and through statistical methodologies, prediction results are obtained regarding an individual's appearance traits, ancestry or age (**Figure 2**) [25].

Inference of traits from genetic markers obtained from samples can provide significant information on the EVCs of an individual and can improve investigative processes. FDP may not yet predict in all appearance traits of an individual with certainty, since some traits can be multifactorial and a result of a complex interplay among different genetic markers and external environmental factors. They can, however, predict the phenotypic traits of an individual at a certain degree of probability that might provide important information to the investigating authorities in narrowing down the number of potential suspects into a smaller group. Unlike with genotype matching-based methods, such as STR profiling, FDP provides evidence i.e. probabilities of characteristics for an unknown individual and it can be used as a part of a wider investigation to identify a perpetrator. It cannot be used so far as a sole evidence in the court, therefore it can only aid police investigations in solving crimes [4]. So far, several studies have already identified and evaluated SNPs that are associated with the prediction of appearance traits such as eye, hair and skin color, hair structure, freckles, baldness and height.

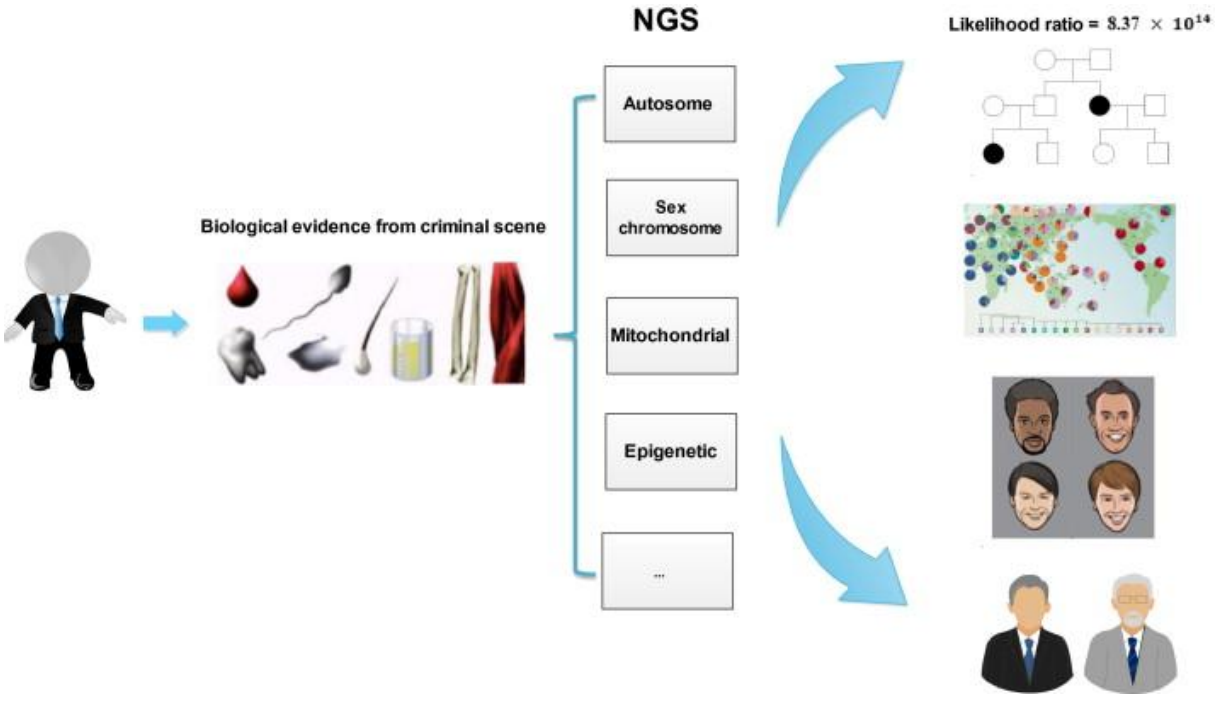


**Figure 2: (adapted from:** [25]**Application of Next-generation Sequencing Technology in Forensic Science. Copyright © 2014 The Authors. Production and hosting by Elsevier B.V.):** Information extracted by next-generation sequencing (NGS) for forensic DNA phenotyping purposes.

## 8.2 Legal and ethical issues around human identification through DNA

Advances in DNA profiling, as well as application of predictive DNA markers in FDP, have on the one hand helped police authorities in solving cases but on the other hand have raised a number of legal and ethical issues and the methodology has been challenged by various interest groups. Most of those objections were raised on legal grounds, motivated by protection of the rights of victims and suspects involved in a case.

The different nature of STR and FDP and as well the different types of data that they use, give rise to different types of ethical concerns. Understanding the way that genetic data are obtained and analyzed as well as the information that those data provide, plays an important role in perceiving the way that those data are used and the real ethical and legal impact of forensic genetics to the society. It is a very sentitive topic and everyone should

be aware of the ethical limitations that surround human identification through DNA. In the present study, I briefly discuss the basic limitations that surround FDP, while for further information and the limitations of STR the literature should be advised [26-30].

### 8.2.1 Legal and ethical issues of FDP

FDP methodology poses a range of ethical and social issues that are referring not only to the nature of information that the data of FDP can provide but also several issues regarding privacy and data protection [26]. Currently within the EU, FDP is not widely used. The only countries where FDP is explicitly regulated are Slovakia and the Netherlands, while for the rest of the countries legislation on use of FDP is ambiguous or absent [31]. More specifically, in the Netherlands, prediction of eye and hair color as well as sex and biogeographic ancestry are allowed and practiced, while in Slovakia testing for "visible phenotypic traits" is also permitted [31]. On the other hand, for other countries such as Germany and Belgium, the current legislation is interpreted as it completely forbids the use of FDP. However for Germany and Switzerland there are ongoing discussions that aim to change the current regulation on FDP [31].

According to the literature, there are eight different ethical and societal issues that are associated with the implementation of FDP [26]. Those are:

1) Discrimination
2) Privacy
3) Data protection
4) FDP as 'biological witness'
5) Stakeholders' inflated reliability, inaccurate test results
6) Cost-benefit of FDP
7) Bias
8) Misuse of the technology

Of those, the most dominant issues appear to be discrimination and privacy. Here I am going to analyze and discuss briefly in an integrated manner the two main issues that are surrounding FDP, especially when predicting appearance traits, age and biogeographical ancestry (BGA). It is important to mention that here only basic and indicative issues regarding FDP are presented, while for further details on the topic the study of Samuel and Prainsack should be advised [26].

#### 8.2.1.1 Discrimination

Discrimination is one of the most prominent ethical and societal issue which is not associated only with FDP, but also in general with the implementation of forensic technologies (e.g. STR profiling) in the criminal justice system. It is referring to the ethnic or religious biases that might result from the use of FDP as an attempt to find the perpetrator of a crime, which as a result could lead to a stigmatization of minority groups in the society, especially when BGA is about to be predicted [32-34].

Discrimination can occur due to the unconscious biases that some people hold, even if they endeavor to 'do the right thing' and act morally. Unconscious biases are beliefs and social stereotypes that have been formed over the

years for certain groups of people or communities and enhance someone's tendency to categorization. Concerns regarding FDP are related to possible unconscious biases of members of law enforcement or criminal justice system that in case that a prediction reflects those beliefs, they might be quicker in accepting the finding than otherwise [26]. Furthermore, the probabilistic nature of FDP outcomes could lead police officers in search for members of the wrong population groups due to false leads, and as a result could lead in accusing innocent individuals.

Another difficult and controversial question is whether FDP outcomes should be public or not. Especially when referring to BGA, which is a very sensitive social topic, caution should be advised on how those outcomes should be communicated to the public, since they might have harmful effects in terms of community relations. One characteristic example is the famous case of 'Phantom of Heilbronn' where a female Polish worker contaminated cotton swabs with her own DNA at the factory where she was working while packaging them. For several years (1993-2009) her DNA was connected to several crimes and burglaries in Germany, France, and Austria and the DNA analysis conducted at that time showed an Eastern European woman as the source of the samples. Due to the fact that the same DNA was connected with a series of crimes that covered a large geographic district, some police officers falsely tended to suspect that the perpetrator belonged to a 'traveler population' such as Roma or Sinti [35].

As a response to the above concerns, the forensic molecular geneticists Manfred Kayser and Peter Schneider argue that similar outcomes and reactions can be obtained in a case where an eye witness connected the suspected perpetrator with a minority group [36]. Therefore the problem does not lie with FDP itself, but rather with the consideration whether or not should we make FDP outcomes public, especially when predicting BGA that can have further consequences [32].

#### 8.2.1.2 Privacy

Issues of privacy have been discussed thoroughly in the literature and have raised questions regarding EVC, age and BGA prediction. Those questions pertain mostly to the outcomes that genetic markers can provide, which often can be used for non-EVC prediction, such as proneness for certain diseases or stigmatizing characteristics i.e. violence or predisposition to homosexuality [37]. In the literature, there is support for the argument that this should not be considered as a privacy data violation as long as the data are not stored in any central police or criminal justice data base. Especially when we are talking about FDP for EVC prediction, we should not consider that any aspect of the suspects' privacy is violated, since those traits are known not only to the person itself but also to all the people who have seen him or her and therefore it cannot be considered as private data [36-40].

However, some scholars are more cautious and argue that we cannot say that this information does not belong to personal/ private data, since it comes from genetic traces. In addition, a very sensitive issue of privacy appears especially in cases where the face of an individual does not reflect his genetic composition. Such examples are when individuals participated in hair dying or plastic surgeries in order to alter some of their visible attributes and when they have done so, they might prefer to keep this information private [41].

Another ethically problematic aspect is a possible BGA prediction through FDP. While BGA can be visible for some people to some extent, especially when there are low levels of admixture, this does not apply in cases of mixed BGA and people who have ancestors from different geographic regions and therefore it can be considered as private information [38]. This information might affect the suspect, especially in the case that the results are not

in accordance with his personal beliefs based on his own cultural or familial identity or it reveals information about his relatives that until that time was unknown[34, 42].

Ultimately, issues are also raised when the traits or BGA to be predicted can provide information about health issues and proneness to certain diseases. In this case, is highly possible that the person is not aware, especially when the disease has not appeared yet or it has not yet revealed its symptoms. Some scholars suggest that this probabilistic health-related information should be refrained from being communicated to the affected individual, while others suggest an approach in which the suspect is asked whether they are willing to be informed about any health-related information that might be revealed throughout the testing. In any case, it is prominently expressed in the literature that the privacy and the right not to know is not universal and different values and rights should be taken into consideration [26].

## 8.3 SNPs as a causal factor for variations in human appearance traits

Humans exhibit a remarkable variety of appearance traits. These include many external features such as eye, hair, skin color, hair structure, freckles, height, nose size, head shape, and brow area [43-45]. Each of these characteristics can be partially explained by genes and partially by other external or environmental factors.

In FDP, the main aim is to predict different appearance traits based on variants throughout the genome. SNPs are the most common type of genetic variation and are results from the substitution of only a single base at a specific position in the DNA sequence [46]. Some variations have been found to be responsible for diversity in humans within and across various population groups. Diversity conferred by SNPs is not limited to appearance traits but extends to other characteristics such as drug response and susceptibility to certain diseases [47]. SNPs can occur either in the coding or in the non-coding regions of the DNA. Protein-coding region or coding sequence is a portion of the genes that encodes protein sequences while non-coding regions do not encode proteins, but can be involved in regulatory processes. SNPs found in the coding region are separated in synonymous and non-synonymous SNPs, and the difference between them is that the non-synonymous affect the encoded amino acid while the synonymous variants do not result in a change of the protein product. SNPs can also affect the gene expression, the messenger RNA degration and the subcellular localization of proteins causing in this way variations associated either with diseases or traits [47]. In the genome, the distribution of the SNPs is not homogenous as most of the SNPs are obtained in the non-coding region of the DNA and natural selection, genetic recombination and mutation rate are factors that affect their density [48]. Furthermore it is obtained that SNPs can differ among the population groups, meaning that certain alleles can be common in some population groups and rare for some others. The minor allele frequency within different populations has been of crucial importance in the identification of population structures and development of ancestry prediction models [49, 50].

In the field of forensic genetics, SNPs were initially used for matching DNA profiles but were quickly replaced by the STRs due to their higher mutation rate[51]. Later on, with the development of the next-generation sequencing technologies (NGS), SNPs were used in forensics for predicting phenotypic traits of an individual such as eye, hair, skin color, and biogeographic ancestry providing potentially significant leads in police investigations by narrowing down a set of suspected perpetrators. This can be used especially in cases where DNA samples are degraded, as SNPs are less prone to degradation, or when there is no genetic match with the conventional method of STR profiling. However, in contrast with the STR profiling, the phenotypic predictions obtained by SNPs are more uncertain, so that the outcomes cannot be used as sole evidence in the court but can only provide aid in police investigations [4].

In this study, I am focused on the use of FDP for predicting a set of appearance traits. Of all traits, only eye, hair and skin color can currently be predicted from crime-scene DNA through forensically validated tools. However, studies have already been conducted in order to identify markers that are correlated with more appearance traits including hair structure, freckles, male pattern baldness, and height [52-57]. Although a set of genetic markers has already been identified for those traits, there are so far no forensically validated tools for their prediction. This is due to the fact that these traits are complex and the phenotypic variations explained by genetic variants remains low. Here I am going to discuss briefly for the predictive models and the genetic markers that have already been identified so far for each one of the traits within the framework of FDP.

### 8.3.1 Genetic variants associated with eye color

Human eye color is one of the traits with high variability, with colors ranging from light shades of blue, intermediate shades such as grey, green and hazel, to dark shades of brown or black. These variabilities are found mostly in people of European descent, and to lesser degree in people from other geographic regions such as Middle East, Asia, or Africa [58]. Brown eye color is considered to be the ancestral human trait and it is found in all the geographic regions in the world, although with a lower frequency in Northern Europe [59]. On the other hand, non-brown eye color is considered to have emerged in Europe first, with a subsequent positive selection [60, 61] possibly as a result of environmental adaptation [61, 62].

Eye color is a quantitative trait that has been shown to be predictable with high accuracy by using only a small set of genetic markers [63]. Several GWAS and linkage studies have revealed genes that are associated with human eye pigmentation [53, 59, 60, 64-71]. The *OCA2* gene on chromosome 15 and more specifically the SNP rs1800407 was initially considered to be the most informative predictor for eye color due to its association with the protein that is required for the processing of the melanosomal proteins [72]. A base change from cytosine (C) to thymine (T) can cause a change from brown to non-brown eyes. However, recent studies showed that another SNP, namely rs12913832 in *HERC2* gene is also strongly associated with human eye color. More specifically, the promoter region for *OCA2* is located in *HERC2* gene, therefore *HERC2* regulates the expression of *OCA2* [73]. The T allele of rs12913832 is likely to co-occur with brown eyes, as opposed to the C allele, which is strongly associated with blue eyes. In addition, deletions in *HERC2* region may cause a decrease in the amount of melanin (hypopigmentation) (**Figure 3**) [73]. Additional work is needed in order to understand the full biological function and interaction between these two genes. Apart from *HERC2* and *OCA2*, SNPs in other genes appear to be involved in human eye color, however with much lower variant effects. Such examples are *SLC24A4* [65, 66], *TYR* [53], *TYRP1* [66], *SLC45A2* [65], *IRF4* [65], *ASIP* [67], *LYST* [74] and *DSCR9* [74]. Of those, six SNPs from six genes, namely *HERC2, OCA2, SLC24A4, SLC45A2, TYR,* and *IRF4*, can predict the eye color with high accuracy, especially for the categories of blue and brown. Based on these genetic variants, IrisPlex, one of the first phenotyping tools was developed and validated. This tool allows the differentiation between blue and brown eyes with accuracy greater than 90% [5, 75], for both homogenous and admixed populations. Regarding the prediction of intermediate eye color, it is lower compared with the other two categories and probably further research is needed in order to identify new genetic variants that contribute to these shades. Thus far, one study from Pospiech et al. [76] showed some gene-gene interactions that are related to the intermediate eye color, from the genes that are mainly related to pigmentation (*HERC2, OCA2* and *TYRP1*), motivating and aiding deeper investigations of future studies on the topic .

Another topic that is discussed in the literature is that if sex is a factor that influences eye color pigmentation. Thus far, some studies have shown that there is an unknown sex-related factor that contributes to human eye pigmentation [77, 78]. More specifically, there is a tendency that males are more likely to have blue eyes than females, who show higher frequencies in prevalences of brown and intermediate eye colors, especially in specific populations. However, further research is required in order to identify the gene-related factors that are responsible for this differentiation between females and males.

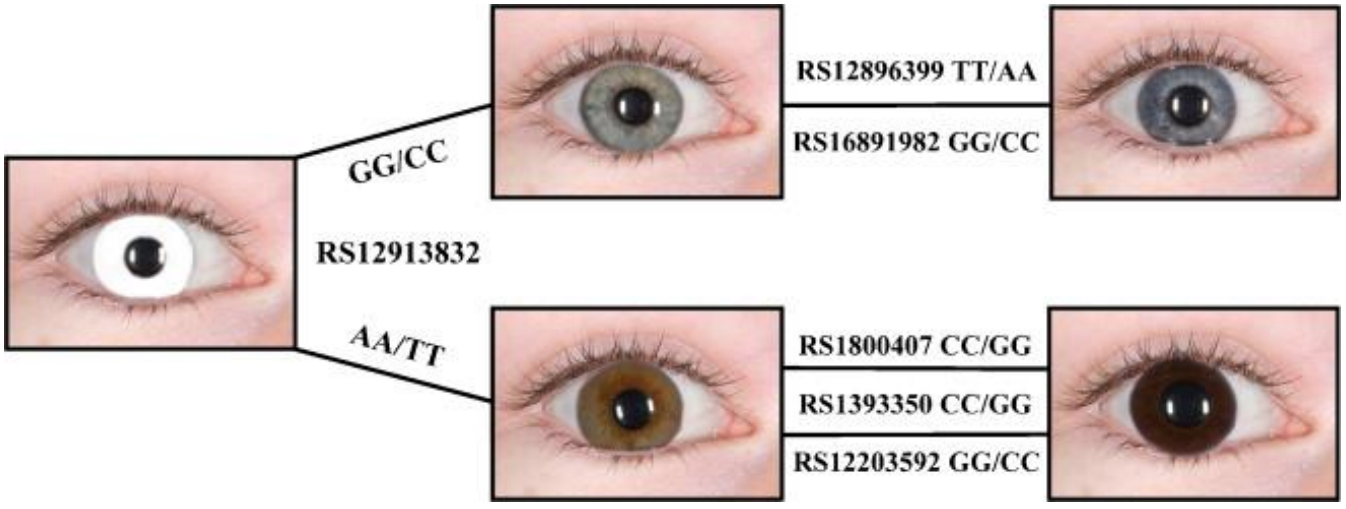


**Figure 3:** Hypothesized scenario for genetic determination of brown and blue eye colors showing the impact of the 6 SNPs include in IrisPlex for eye color prediction. Adapted from:[5]

### 8.3.2 Genetic variants associated with hair color

Hair color, similarly as eye color, is a trait with a wide range of phenotypes, especially in people of European descent and nearby regions, such as Western Asia and Middle East [58]. People from other geographic regions exhibit predominantly black hair color, which is considered the ancestral phenotype together with brown eye color.

Variations in hair pigmentation, similarly as in eye pigmentation, are considered to be influenced by sexual selection [61]. The main differences among the various hair color categories are the result of expression of two types of melanin: the brown/black eumelanin and the red/yellow pheomelanin [79, 80]. These are the two different types of melanin synthesized by the melanocytes, and their ratio determines hair pigmentation. Eumelanin is a dark pigment that is responsible for brown and black color and therefore is predominant in people that have this pigmentation. There are two types of eumelanin; brown and black. The brown eumelanin in small amounts without co-occurrence with other pigments leads to blond hair color, while large amounts of brown eumelanin or black eumelanin results in brown and black hair colors respectively. On the other hand,

pheomelanin is a lighter pigment that impacts red and yellowish hair color. In cases where small amount of brown eumelanin is mixed with red pheomelanin, the phenotypic result is red hair color. Pheomelanin is expressed in the redder areas of the skin as well, such as human lips and people with higher expression of pheomelanin cannot produce high amounts of eumelanin. On the other hand, people with dark hair color may still produce the lighter pigmented pheomelanin, but it might be covered by the darker eumelanin and therefore is not or only subtly visible, resulting in auburn tones of brown hair. The color categories influenced by different types of melanin are defined, at least within the forensic framework, as follows: blond, brown, red, and black. So far, several genes and their SNPs have been identified by various studies for their contribution in the expression of different hair pigments [53, 65-68, 81-85]. One example is the *MC1R* gene, which is primarily responsible for the melanin produced in the human body, and its activation can cause the melanocyte to produce pheomelanin instead of eumelanin. Its activation is also associated with other human appearance traits, such as freckles and fair skin color, and diseases such as skin cancer. Its high accuracy for predicting red hair color (84%) has been known for about twenty years ago [86].

Later, other genes responsible for hair color variation were identified and the first predictive model was developed by Branicki et al. based on 13 SNPs [87]. This model reached an overall accuracy that ranged from 81% to 93% for each hair color category. More specifically the accuracy was 81% for blond, 82% for brown, 87% for black and 93% for red hair color. In 2013, another predictive tool was developed for hair and eye color. This tool, called HIrisPlex, comprises 22 SNPs from 11 genes, namely *MC1R, HERC2, OCA2, SLC24A4, SLC45A2, IRF4, EXOC2, TRYP1, TYR, KITLG,* and *ASIP* and includes the six markers that were already included in IrisPlex [6]. This model reached an overall accuracy for hair color prediction ranging from 75% to 92%, similar to the model already developed by Branicki et al.[87]. One difference between those two models was that the first one was trained in one single country in Eastern Europe, namely Poland, while HIrisPlex was trained in a dataset with more diverse phenotypes that contained DNA samples from three different European countries, namely Ireland, Greece and Poland. From the aforementioned genes, the *TYRP1* was identified by Kenny et al. [88] to contribute to the blond hair pigmentation in Melanesian individuals. More specifically, its mutation in the 93C allele encodes an enzyme for melanin biosynthesis and is therefore related to pigmentation. However, this specific mutation does not contribute equally when it comes to the blond hair phenotype in Europeans [88].

One issue that the current hair prediction models are facing and is discussed in the literature is the age-related hair color darkening. This phenomenon is referring to the changes of the hair color from childhood to adulthood. More often, children tend to have lighter hair color at younger age, which gradually changes to darker shades when they are growing up. Most prediction models were trained on phenotypic information of adults, therefore they are not taking into account possible informative markers that are responsible for this phenomenon. That partially explains the lower prediction accuracy for blond hair, compared with the rest of the categories. Only one study on the subject was conducted including individuals from 6 to 13 years old and showed that HIrisPlex fails to correctly predict the trait in individuals that were blond at younger age [89]. This indicates the need to identify more markers in order to improve the current predictions for hair color, especially for blond hair, which shows the lowest prediction accuracy compared with other categories.

### 8.3.3 Genetic variants associated with skin color

Skin color is a trait that, similarly with the eye and hair color, has a wide range of phenotypes and it is considered to be an adaptive and a genetically complex trait. Its phenotypic variations are a consequence of the quantity and size of the melanin particles that are located in the epidermis and are produced by the melanocytes [90]. The two aforementioned types of melanin, namely eumelanin and pheomelanin, also appear in this part of the human body, with the first type, eumelanin to be expressed in individuals with darker skin tones, while pheomelanin to be expressed in people with lighter skin tones. It is known that differentiations in skin color are affected by environmental factors as well such as age, drugs, diseases, or the levels of exposure to UV radiation [91]. Although the phenotype is variable within the lifetime of an individual, it is hypothesized that UV exposure was one of the environmental factors driving inborn pigmentation background in humans. Individuals from regions close to the equator, where UV radiation is higher, tend to have darker shades of pigmentation while in more distant regions where UV is less intense, fairer skin tones are dominant. Melanin works as a protector and tends to absorb UV radiation, which in high levels could cause DNA damage and health problems such as skin cancer [92].

In the context of forensic genetics, there are five established categories for skin color prediction, namely very pale, pale, intermediate, dark and dark to black. Despite the current lack of consensus regarding external factors contributing to skin pigmentation, there are several studies that focused on skin color prediction based on the most strongly associated SNPs. Early studies started by analyzing skin color variation in single homogenous population groups, such as the study of Stokowski et al. that included South Asians [93] and the study of Jacobs et al. that was focused on Europeans [94]. Later on, other studies were conducted that tried to identify associations between genotypes and phenotypes either in admixed or in homogenous populations. The result was, that associations were identified in admixed individuals, but they appeared to be less discriminative for more homogenous populations [95, 96].

More recently, in 2018 the first predictive tool for skin color was developed and forensically validated. This tool, called HIrisPlex-S was an extension of the aforementioned tools IrisPlex and HIrisPlex for eye and hair color prediction [7]. It comprises 36 SNPs in total, located in 16 genes such as *SLC24A5, HERC2, SLC45A2, KITLG*, and *IRF4.* In addition, it includes the 22 SNPs that were previously used for eye and hair color prediction. HIrisPlex-S can predict the five previously mentioned skin color categories with accuracies ranging from 72% to 97%, and together with IrisPlex and HIrisPlex are publically available at https://hirisplex.erasmusmc.nl/.

Similarly to hair color, the *MC1R* gene appears to be responsible for the skin color, since its mutations are affecting the production of a certain type of melanin in the human body [97]. Other examples are the *KITLG* and the *ASIP* genes with mutations that are associated with lighter skin colors in European and Asian populations [98].

Despite the fact that skin color can be predicted with high accuracy, as with hair and eye color, more studies are needed in order to identify more genetic variants or environmental influences that can possibly improve the already existing approaches and provide an insight of their mechanisms and interactions.

### 8.3.4 Genetic variants associated with hair structure

Hair structure or hair shape variation is a visible trait that shows a strong diversity among different continental groups. More specifically, non-straight hair is more dominant in individuals of African origin and less dominant in Europeans. On the other hand, Asian populations show high prevalence of straight hair [99]. Differentiations in

hair structure are also obtained within the continental groups, where people from the same group can have different levels of curliness. It is considered a high heritable trait and the heritability of curly hair in Europeans is estimated to be up to 95% [100]. As previous studies showed, hair shape variability can be classified into eight main groups, starting from classic to more sophisticated verbal descriptions, with terms such as straight, wavy, curly, frizzy, woolly, kinky, helical among others [101]. However, in the context of forensic genetics, where precision is necessary, the above scale could lead to confusions because of the unclear definitions and limits of each category. For this reason, more simplified scales are used and mostly the one that categorizes hair shape in three types namely straight, wavy and curly [102].

Regarding the genetic basis of hair structure, several genome-wide association studies (GWASs) have already identified eight genes that are responsible for the differentiations in human hair shape among different continental groups. These are *TCHH* [103-105], *EDAR* [105-107], *GATA3* [105], *PRSS53* [105], *WNT10A* [103, 104], *FRAS1* [103, 104], *OFCC1* [103] and *LCE3E* [103]. From them, variants of the *TCHH* gene are identified to be associated with hair straightness in Europeans while for the same phenotype in other populations, other genes might be responsible. One example of the above is the *EDAR* gene which is the predominant gene for straight hair and hair thickness in Asian populations [106]. This gene interacts with a protein called ectodysplasin A1 and they trigger a series of chemical signals that affect several cell activities such as growth [108]. Other genetic variants such as the ones found at *GATA3* and *PRSS53* genes are highly expressed in the hair follicle and were found to affect the hair structure mostly in Native Americans and in Latin Americans or mixed Europeans [105].

These genetic variants explain only a small proportion of the hair shape variation in humans and only few recent studies so far have been conducted in order to develop a predictive model for this trait. One of the first studies was the one of Pospiech et al. in 2015, where three different models were compared and evaluated (logistic regression, regression and classification trees and neural networks) [109]. This model was based on six SNPs located in *TCHH*, *FRAS1* and *WNT10A* genes and the dataset used comprised of samples of Polish origin. Later on, in 2018 Pospiech et al. developed another model based on an extended set of 90 SNPs and a dataset containing individuals from Europe, Asia and admixed Europeans and Asians [102]. The model included additionally sex and age as predictive factors, which slightly improved the overall accuracy, however more studies are necessary in order to identify more markers that contribute to hair structure. In this way we might improve the predictive accuracy of the already existing models for hair structure prediction and allow a predictive tool to be established and used in FDP expanding the current set of the FDP models beyond pigmentation traits.

### 8.3.5 Genetic variants associated with freckles

Freckles or ephelides is an appearance trait that is observed on the skin surface as hyperpigmented spots. It mostly appears in European and Asian populations and especially to those who have fair skin color and red hair. Typically they appear in childhood but they can increase or disappear in adolescence [110, 111]. Despite the fact that this appearance trait is affected by the exposure to UV radiation, its occurrence has a strong genetic background. So far, several genes and their variants have been established for their association with freckles that overlap with other pigmentation traits [53, 66, 103] or are associated with skin cancers as well [112-115]. These genes that have already been identified include *MC1R*, *IRF4*, *TYR*, *ASIP*, *OCA2*, and *BNC2* [53, 66, 103]. Out of those, the *MC1R* provides the major contribution in freckles occurrence, especially for individuals of European origin [116, 117] and its variants contribute to a more severe phenotype that includes fair skin, red hair and

freckles, the so called RHC phenotype. These *MC1R* variants include D84E, R151C, R142H, R160W, D294H, and I155T [83, 117] and they are known as R alleles, since they have high penetrance and they restrain the function of *MC1R*. Other variants that have lower penetrance are known as r alleles, while the ones that do not have a significant effect on *MC1R* function are called pseudoalleles [118].

The first study for freckles prediction was conducted by Hernando et al. in 2018 [119] and it was based on the variation in four pigmentation genes adjusted for sex and the additional information of whether the individuals obtained freckles during childhood or adolescence. In addition, information on age and pigmentation traits was included in the study. The dataset comprised of individuals of Spanish origin and the model was based on multivariate logistic regression. A model selection was then conducted according to the lowest Akaike Information Criterion (AIC) in order to find the appropriate marker set. The analysis showed that a prediction model for freckles can be built based on five genetic predictors, namely R and r variants in *MC1R, IRF4* rs12203592, *ASIP* rs4911442 and *BNC2* rs2153271. The model was able to predict freckles incidence with a sensitivity up to 60% and an area under curve (AUC) equal to 78% for two categories, non-freckled and freckled.

A recent study from Kukla-Bartoszek in 2019 revealed 19 DNA variants that are associated with the freckles phenotype and additionally 12 independently contributing predictors [120]. Two different models were developed and compared for two different category scales. The first one was a simplified binomial logistic regression model with 12 predictive variables that classifies individuals in two categories, freckled and non-freckled. The second one was a multinomial logistic regression model based on 14 predictors and the individuals are categorized into three categories, non-freckled, medium freckled and heavily freckled. In this case, the categories of non-freckled and freckled were predicted with an AUC value of 75% and 79%, respectively, while the medium freckled phenotype appeared to be more complex reaching a level of 65% of AUC.

Of course, there is a need for greater understanding of the genetic basis of freckles especially for building a predictive tool for forensic investigations where prediction accuracy is of utmost importance. However, the current findings and the prediction accuracies achieved so far should not be underestimated. The number of correct predictions could possibly be improved by setting a threshold when interpreting the predictions. Furthermore, additional studies, genome-wide analyses or larger cohorts of individuals could possibly help in expanding the already existing marker sets and identify those genes that explain the variability of this trait.

### 8.3.6 Genetic variants associated with male pattern baldness

Androgenic alopecia, also known as male pattern baldness (MPB) is a trait that is more frequently observed among men of European origin. It affects approximately 20% of men aged 20 and increases steadily by age reaching 90% for men around 90 years old [121]. Is a trait that has not only alterations in the physical appearance but also has substantial effects on psychological functioning and social processes [122]. Other studies showed that MPB is also associated with certain diseases such as cardiovascular diseases [123-125] and prostate cancer [126-128]. Whilst the etiology of MPB is still not fully understood, current studies show that genetic proneness and hormonal dependence is behind the occurrence of this appearance trait. More specifically, around 80% of the variability of MPB can be explained by genetic factors [121, 129]. One gene that is found to be associated with MPB is the *HDAC9* gene, which is expressed in hair follicles [130]. It can have a direct or indirect interaction with the AR protein in the *AR* gene and plays a role in the regulation of the *AR* gene and subsequently can affect the phenotypic expression of the MPB [130, 131].

In the study of Marcinska et al. a set of 50 SNPs for MPB were analyzed in order to identify their predictive ability for this trait [132]. MPB prediction was based on the Norwood-Hamilton scale which contains eight main categories (**Table 1**) and identified markers with the major contribution to MPB were among those in region q12 on X chromosome, on 20p11 and in the genes *HDAC3, EBF1* and *TARDBP*. From them, the first predictive model was made, containing the following SNPs: rs5929324 near *AR*, rs1998076 in the 20p11 region, rs756853 in *HDAC9*, rs929626 in *EBF1* and rs12565727 in *TARDBP*. This model reached an overall accuracy expressed in AUC equal to 76% and with threshold value of 50%, the number of corrected predictions was equal to 66%. When increasing the threshold value to 65% then the number of corrected predictions was raised up to 75.8%.

The second predictive model was an extension of the first one and included the aforementioned five SNPs with the major association in MPB and additionally 15 SNPs, namely rs1041668, rs6625163, rs6625150, rs962458, rs12007229, rs2180439, rs913063, rs1160312, rs6113491, rs6461387, rs6945541, rs7349332, rs4679955, rs9668810, and rs10502861 [132]. In this case, the AUC was raised up to 86%, but for the age group of 50 years and older. That demonstrated that when combining lower prediction markers with the ones that have a stronger association, the prediction accuracy can be improved. However, further studies are necessary in order to identify more markers that possibly contribute to MPB and also more information is needed on the genetic background of the senescent alopecia.

**Table 1 Simplified phenotypic description of the Norwood-Hamilton baldness categories [133]**

| **The Norwood-Hamilton scale of male-pattern baldness** | **Phenotypic description** |
|---|---|
| Grade I | No recession of the hair line |
| Grade II | Minor recession of the frontal hairline |
| Grade III | Deep symmetrical recession at the temples |
| Grade III vertex | Significant frontal hair loss coupled with hair loss at the crown of the head |
| Grade IV | Deepening frontal recession in the temples and progressively more hair loss at the crown |
| Grade V | Hair loss at the vertex and front temporal areas are extended |
| Grade VI | The frontal and vertex regions of hair loss merge into one area and increase in size |
| Grade VII | The most advanced stage of male-pattern baldness, in which all hair is lost along the front hairline and crown |

### 8.3.7 Genetic variants associated with height

Adult height is an appearance trait that has been studied extensively, since is helpful not only for forensic investigations but also for several areas including pediatric endocrinology [134]. Four GWAS studies focused on tall stature conducted by the Genetic Investigation of Anthropometric Traits (GIANT), demonstrated that this trait has a very high genetic complexity, meaning that many hundreds and probably thousands of independent genetic loci are contributing, which can be characterized by several SNPs that have a small effect on height [56, 135-137]. It is considered to be a highly heritable trait, with heritability estimated up to 80% [138-140].

Currently there is no forensically validated tool for predicting height; however studies have identified a set of the SNPs that explain a certain variance of the tall stature. In 2013, a study conducted by Liu et al. showed a prediction model containing 180 markers associated with human height [134]. Those 180 genetic markers were previously identified by the first GIANT study on Europeans [141]. In this model the prediction accuracy, expressed by AUC, was equal to 75% for binary classification of tall stature.

In the second GIANT study, an extended set of height-associated SNPs was identified. More specifically, 697 variants were identified that explain one-fifth of the heritability of the adult height [136]. Furthermore, it was shown that ~2000, ~3700 and ~9500 associated SNPs explain only ~21%, ~24% and ~29% of height variability, respectively, highlighting the complexity of the trait, with each SNP explaining only a small fraction of its variability. Those 697 genetic markers were used in the more recent study of Liu et al. in 2019 in order to predict human height in a set of Dutch Europeans [141]. In this study, quantitative analysis and logistic regression models were applied for binary height prediction. The results of this study showed that a fairly accurate prediction was achieved, with AUC equal to 79%. Furthermore, an additional model with a reduced number of SNPs, namely 412, was also presented that showed a slightly inferior AUC equal to 76%. Despite the slight improvement obtained in the full model, tall stature prediction still remains a challenging task, especially in FDP. In order to improve the prediction accuracy for adult height, probably more independent SNPs needed to be identified and also identification and incorporation of possible external or environmental factors that contribute to height are necessary, especially in non-European populations.

## 8.4 Statistical classification problem and its applications

Statistical classification is referring to the problem of training a model in order to separate new observations into classes (or categories) according to their common characteristics. It is a problem of either supervised or unsupervised learning. Supervised learning means that it determines the predictive model by using data points with already known outcomes. In other words, the model is trained not only by the input data but also with the correct output data [142]. On the other hand, unsupervised learning is performed with unlabeled data and the model tries to detect the unknown, hidden structure of the data [143]. The history of classification starts back at 1940s where discriminant analysis was developed based on the multivariate normal distribution [144]. Later on, with the increase of computing power, more so-called machine learning (ML) approaches came up and classification became understood in a larger context of learning algorithms. ML algorithms are mathematical models that have the ability to recognize patterns and features in the datasets and use this information for prediction [145]. In general, the more data is available for training, the more accurate the predictions of ML models become. Today, ML is applied in a wide range of applications such as bioinformatics, education, and robotics. Some examples of the ML approaches include linear and logistic regression, decision trees, random forests, k-nearest neighbors, and artificial neural networks.

Some of the ML classifiers have already been applied in forensic science. However, regarding EVC prediction, which is the main focus of this study, the majority of the models were based on the logistic regression. Throughout this study, I am going to apply other methods for predicting appearance traits, including support vector machines, random forests and artificial neural networks, and I will compare their performances with the standard MLR. Here I discuss briefly each of the methods.

### 8.4.1 Multinomial logistic regression

MLR is probably the most widely used method for classification, especially in EVC prediction. The majority of the available and forensic validated tools are based on MLR [5-7]. It is used to predict the probabilities of the different categorical outcomes of a dependent variable based on a set of independent variables. The number of categories of the dependent variable can be two or more. In case of two categories, binomial logistic regression is applied, while for more than two categories its extension to MLR is used. MLR maximizes the likelihood in order to evaluate the probability of each category. Unlike other methods, MLR does not assume normality, linearity or homoscedasticity and it does not require hyperparameter tuning. The MLR model for 3 categories is defined as follows [63]:

$$\ln\left(\frac{p_2}{p_1}\right) = \alpha_2 + \sum_{j=1}^{k} \beta_2 (\pi_2)_j x_j \qquad (1)$$

$$\ln\left(\frac{p_3}{p_1}\right) = \alpha_3 + \sum_{j=1}^{k} \beta_3 (\pi_3)_j x_j \qquad (2)$$

Where $\alpha_i$,$\beta_i$ $(i = 2,3)$ are the regression coefficients and $p_i$ $(i = 1,2,3)$ are the probabilities for each observation that belong to a category *j*. These probabilities are defined as:

$$p_2 = \frac{exp\left(\alpha_2 + \sum_{j=1}^{k} \beta_2 (\pi_2)_j x_j\right)}{1 + exp\left(\alpha_2 + \sum_{j=1}^{k} \beta_2 (\pi_2)_j x_j\right) + exp\left(\alpha_3 + \sum_{j=1}^{k} \beta_3 (\pi_3)_j x_j\right)} \qquad (3)$$

$$p_3 = \frac{exp\left(\alpha_3 + \sum_{j=1}^{k} \beta_3(\pi_3)_j x_j\right)}{1 + exp\left(\alpha_3 + \sum_{j=1}^{k} \beta_3(\pi_3)_j x_j\right) + exp\left(\alpha_2 + \sum_{j=1}^{k} \beta_2(\pi_2)_j x_j\right)} \quad (4)$$

$$p_1 = 1 - p_2 - p_3 \quad (5)$$

Here, $x_j$ are the values of the predictive variables used in the model and *j* is an index for the number of the used input variables. Observations are classified to the category which maximizes the probability $p_i$ obtained.

### 8.4.2 Support vector machines

SVM is a supervised machine learning approach which was devised by Vapnik et al. [146, 147] and is used for classification problems. It can handle two or more outcome classes and its basic idea is to separate the data into classes by finding the optimal hyperplane in the space of input features. The optimal hyperplane is the one that has the maximum distance between the data of different classes and its dimensions always depends on the number of classes. SVM can transform from the original feature space to non-linear feature space in order to simplify the computation. Kernel functions that can be used include the linear, non-linear, polynomial, sigmoid and radial basis function (RBF). RBF is the most widely used and is defined as follows:

$$K(X_1, X_2) = exp\left(-\gamma \|X_1 - X_2\|^2\right) \quad (6)$$

where $\|X_1 - X_2\|$ is the Euclidean distance between the data points $X_1, X_2$ and $\gamma$ denotes the shape and smoothing of the hyperplane. In order to obtain the optimal performance of an SVM, hyperparameter tuning is required and the hyperparameters that are needed to be tuned differ according to the kernel function that is applied. In case of RBF, the hyperparameters that need tuning are $\gamma$ and C, where C indicates the cost of misclassification of the observations. In SVM, the hyperplanes are decision boundaries for classification of the data points. From the data points, the support vectors are formed and the ones that are closer to the hyperplane define its optimal position.

### 8.4.3 Random forests

RF is a method that can be applied to classification problems. It was first proposed by Breiman and it combines classification trees and bagging [148]. It is an extension of the decision trees method and it consists of a number of individual trees that operate as an ensemble. Decision trees is a predictive model that consists of nodes and branches and is expressed as a recursive partition of the feature space to subspaces that constitute a basis for prediction [149]. More specifically, in each node, one feature of the data is evaluated for maximizing the split of the observations during the training procedure. Despite the advantages of the decision trees, such as their easy interpretability, one of its disadvantages is the large variance due to its strong dependence on the given dataset each time. Slight changes in the dataset can significantly affect the performance of a decision tree. This was a main motivation for the development of RF. In RF, each of the trees makes a class prediction for the random samples of the training data, i.e. bootstrapped samples, and in the end the class with the majority of the votes is chosen as the final prediction. Together, trees can provide ensemble predictions that can be more accurate than the individual tree predictions. The basic idea behind RF is that a set of low correlated classifiers can build up a stronger classifier with lower variance [150, 151].

In order to increase the predictive power of the RF there is a set of hyperparameters that needs to be tuned. Hyperparameters include the number of trees, the number of features at each split, the weight assigned to each class, and maximum number of leaf nodes. The first two are considered the most important, however this always depends on the problem and the available dataset. Regarding the number of trees, several studies have been conducted in order to find an optimal number, which always depends on the problem and the dataset. One of these studies by Liaw and Wiener, which argued that the more trees we add, the more stable the results of variable importance are [152]. However, in contrast to this study, others claimed that the large numbers do not always improve prediction results and instead a smaller number of trees can be sufficient. One example is the study of Oshiro et al. that applied RF in 29 different datasets and tried different number of trees [153]. They found out that a range of trees from 64 to 128 can provide good results and additional number of trees did not significantly improve the prediction outcomes. For the optimal number of features at each split, the conventional wisdom states that selecting bigger numbers can increase the strength of the individual trees, but on the other hand, when reducing this number could lead to lower correlation between the trees, which as a result can increase the predicting power of the RF as a whole [154]. Suggested values for classification are $\sqrt{p}$ or $\log p$, where $p$ denotes the number of predictive variables included in the model [154]. However, due to the fact that these values can vary depending on the problem, they should be considered and treated as tuning hyperparameters.

### 8.4.4 Artificial neural networks

ANN are computational models that are designed to simulate the functioning of the human brain. Their history started in the late nineteenth and early twentieth century and since that time several approaches have been developed [155]. ANN can handle a wide range of problems and are widely applied in several fields such as pattern recognition [156], quantum chemistry [157] and finance [158]. The basic idea behind ANN is that, just as the human brain, they consist of interconnected units, the so-called artificial neurons. These neurons are arranged in one or more layers, they are processing the information and their combined effect is activating neurons in a subsequent layer, if existing. The input data are transferred through the network in the forward direction to the neurons in other layers, until an outcome is obtained. Each node has its own weight that is adjusted during training.

Although ANN form an approach that is able to be applied in several problems and can handle various types of data, finding its optimal hyperparameters that can maximize its performance can be sometimes challenging. This is because there is no explicit method for selecting hyperparameters and the optimal ANN architecture depends on the problem. In general, there is a group of hyperparameters that can be tuned that are either related to the network structure, such as the number of hidden layers and nodes, or they can be related to the algorithm training, such as the number of epochs, which is the number of complete passes through the dataset. In any case, finding the optimal hyperparameters for an ANN can be challenging, however it can significantly improve the model performance.

## 8.5 Bayesian classification

Bayesian classification is a statistical method for classification that is based on the Bayesian theorem. The main principle behind Bayesian modelling is that prior probabilities are used in order to represent uncertainties or prior beliefs. Combination of prior information with the data-based likelihoods results the posterior probabilities. The basic idea behind Bayesian classification is [159, 160]:

$$data-dependent\ likelihoods \times prior\ probabilities = posterior\ probabilities \quad (7)$$

The use of prior probabilities is fundamental for the Bayesian approaches. It represents all information or subjective beliefs about unknown parameters before any data are taken into account. Priors can be obtained either from outcomes of previous experiments or assumptions. In case no previous information on a parameter is available, priors can be non-informative, meaning that they can express the lack of available information, but in a principled way. By including priors to our prediction model we obtain as a result the posterior probabilities, which are the revised, or updated, probabilities that occur after considering new information. The data-dependent likelihoods represent the probability that an observation belongs to a certain category.

In classification problems, it is necessary to evaluate the cost of the different types of errors in order to make a rational decision on the category prediction. Similarly as in its fundamental idea, the decision making in Bayesian classification combines both priors and likelihoods in order to achieve the minimum probability of error. If we have a probabilistic variable ω for the different classes, a variable $c$ representing the number of classes and $a_i$ denoting the predictions, then the loss function is defined as follows [161, 162]:

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad (8)$$

where $i,j = 1,\ldots, c$. This is the so-called zero-one loss function and quantifies the cost of a prediction. Hence, the Bayesian risk can be defined accordingly:

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)\, P(\omega_j \mid x) \quad (9)$$

If we consider a decision rule $a(x)$ then the overall risk can be defined as:

$$R = \oint R(a(x) \mid x) P(x)\, dx \quad (10)$$

We are looking for the rule $a(x)$ that minimizes $R(a(x)|x)$ for all $x$, therefore:

$$a = \arg\min R(a_i \mid x)$$

$$= \arg\min \sum_{j=1}^{c} \lambda(\alpha_\iota \mid \omega_j) P(\omega_j \mid x) \quad (11)$$

With the calculation of the posterior probabilities, a decision rule is necessary in order to classify the observations into the different categories. In our case, this is being made according to the maximum posterior probability. That means that for the 3-class problem, an observation is classified according to the maximum posterior $\pi_i,\ i = 1,2,3$. Namely, an observation $\alpha$ is classified according to:

$$\max\{\pi_1,\pi_2,\pi_3\} \rightarrow \{1,2,3\} \quad (12)$$

In FDP, prior information can be extracted from the trait prevalences within the population groups. Hence, inferring the biogeographic ancestry might be helpful in order to identify these prevalence values, which incorporation in the prediction model may improve the already existing prediction accuracies. Here, I

incorporated prior information on the standard MLR model and in this case the posterior probabilities for the 3-class problem will be defined as follows:

$$\pi_2 = \frac{\exp\left(\ln\left(\frac{p_2}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right)}{1 + \exp\left(\ln\left(\frac{p_2}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right) + \exp\left(\ln\left(\frac{p_2}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right)} \quad (13)$$

$$\pi_3 = \frac{\exp\left(\ln\left(\frac{p_3}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right)}{1 + \exp\left(\ln\left(\frac{p_3}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right) + \exp\left(\ln\left(\frac{p_3}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right)} \quad (14)$$

$$\pi_1 = 1 - \pi_2 - \pi_3 \quad (15)$$

where $p_i$ $(i = 1,2,3)$ are the prior probabilities for each of the three categories and $\alpha_i$ and $\beta_i$ $(i = 2,3)$ are the model coefficients. In a similar way, this model can be extended to more than 3 categories and each observation is classified to the category that yielded the higher posterior probability, as previously mentioned in (12).

Of the already established tools and the studies conducted so far for EVC prediction in FDP, the majority of them are based on MLR, with only few exceptions that used alternative methods, while prior information into the prediction modelling was barely used. Some of these examples are the Snipper [163] and the models developed by Maroñas et al. [164] and Söchtig et al. [165] for eye, skin and hair color prediction, respectively, which were based on naïve Bayesian likelihood classification. Among them, the Snipper is the only model that makes use of prior information, in particular. It is the so-called LOCPRIOR information, which refers to the distribution of specific characteristics within a population group and their frequencies in the populations. Such characteristics can be linguistic, geographical, or phenotypic information among others. Other alternative approaches are the 7-Plex and 8-Plex [166] for eye and skin color prediction which were based on binomial proportion tests as well as the classification tree approaches developed by Allwood et al. [167].

## 8.6 Performance evaluation with standard metrics

Evaluation metrics are measurements that help us quantify the performance of different statistical or machine learning models. There are several metrics available in order to test whether a model operates correctly and these include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under curve (AUC), confusion matrix, and overall accuracy among others. Here I am going to briefly discuss the definitions of each of those metrics that are used throughout the study.

A confusion matrix defines basic quantities that are relevant for performance evaluation of a trained model performed on a test dataset, and from which entries performance measurements are derived (**Table 2**). Sensitivity, or true positive rate, is a measurement that refers to the proportion of the actual positive samples that are correctly identified by the model, while specificity, or true negative rate, is the proportion of the actual

negative samples that were correctly classified by the model. PPV, or precision, measures the proportion of the correct classifications among all predictions of the category tested, while the NPV measures the proportion of the correct classifications in all predictions other than the category that is being tested (**Table 2**). Area under curve (AUC) is a performance measurement for classification problems that quantifies the ability of the model to separate the observations into the different classes. Ultimately, the overall prediction accuracy is the ratio of the correct predictions to the total number of input samples.

**Table 2: Table demonstrating the derivation of sensitivity, specificity, positive and negative predictive value from a confusion matrix**

| | | Truth: Has the trait | Truth: Does not have the trait | |
|---|---|---|---|---|
| Test | Positive | True Positives (TP) | False Positives (FP) | $PPV = \frac{TP}{TP + FP}$ |
| | Negative | False Negatives (FN) | True Negatives (TN) | $NPV = \frac{TN}{TN + FN}$ |
| | | **Sensitivity** $\frac{TP}{TP + FN}$ | **Specificity** $\frac{TN}{TN + FP}$ | |

# 9 Aims of the PhD Thesis

Prediction of EVCs through DNA has become a topic of major focus in FDP in the last decades. So far, there are established and forensically validated tools that focus on the prediction of traits such as eye, hair, and skin color. Available tools can predict these traits with high accuracy from a relatively small number of genetic markers, especially for eye and hair color. For the rest of the traits, there are currently fewer studies focusing on their prediction due to their genetic complexity and the limited knowledge on the genetic markers or the external/environmental factors that contribute to the phenotypic variations of these traits.

Motivated by the current state and the shortcomings in EVC prediction, the aim of my project was mainly the testing of the impact of prior information for an extended set of EVCs including not only eye, hair and skin color, but also hair structure and freckles. In our case, prior information indicates the prevalence data for each trait category among different population groups. Furthermore, a comparison of four different machine learning (ML) classification methods was conducted for eye, hair, and skin color, in order to see whether any of them outperforms the standard MLR method.

Thus, in the first study I aimed at the compilation of the prevalence values, which later on were supposed to be incorporated in a Bayesian MLR model for EVC prediction. During the literature review on trait prevalences, I found out that the amount of available and reliable data was quite limited. Therefore, I managed to collect trustworthy and population-representative data only for eye and hair color for specific population groups, mostly European. Given the lack of accurate prevalence data, their incorporation into the EVC prediction was not feasible at this stage. Subsequently, in order to assess the impact of priors I proceeded in the second study with a prior-based prediction model, which considers a grid in the complete space of all possible prior values for each trait category, checking in this way the possible effects, including also cases of prior misspecification. The appearance traits that this model was applied to included eye, hair, skin color, hair structure, and freckles. The prior incorporated model was based on the already reported marker sets from previous studies [5-7] for each trait and its performance was compared with the prior-free approach.

In the third study, I aimed at the evaluation and the comparison of four different ML classification methods. ML is a widely used set of methods applied in a wide range of fields. In the context of forensic genetics, one ML method that is mostly used so far, especially for EVC prediction, is the MLR, while other ML methods such as support vector machines (SVM), random forests (RF) and artificial neural networks (ANN) have not been thoroughly applied in this field. Therefore, I aimed at a systematic quantitative comparative analysis between the standard method of MLR and the three aforementioned ML methods, namely SVM, RF, and ANN. The traits that I covered here were eye, hair, and skin color. All models were compared according to the standard performance measurements, namely overall accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under curve (AUC) in order to evaluate whether the alternative ML methods outperform the conventional method of MLR for DNA-based EVC prediction.

# 10 Major findings

## 10.1 Published results

- Available trait prevalence information is restricted for eye and hair color and is available mostly for European populations.
- Moderate associations between eye and hair color.
- Lighter pigmentation dominates in Northern Europe while darker pigmentation dominates in Eastern countries.
- Lack of available and reliable data makes incorporation of spatial trait prevalence as prior knowledge in EVC prediction not feasible at this stage.
- A proportion of priors shows potential to improve EVC prediction.
- Misspecification of priors can significantly deteriorate the model's performance.
- Impact of priors is likely inversely related to genetic determination.
- None of the three ML methods outperformed MLR, at least with the currently available set of predictive markers.
- No significant differences among the four classifiers might be due to the nature of the traits or due to unknown factors.

# 11 Published main investigations

## 11.1 True colors: A literature review on the spatial distribution of eye and hair pigmentation

This text was published as an article in Forensic Science International Genetics in 2019, 39; 109-118, https://doi.org/10.1016/j.fsigen.2019.01.001,

Forensic Science International: Genetics 39 (2019) 109–118

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

# True colors: A literature review on the spatial distribution of eye and hair pigmentation

Maria-Alexandra Katsara, Michael Nothnagel*

Cologne Center for Genomics, University of Cologne, Cologne, Germany

ARTICLE INFO

Keywords:
Eye and hair color
Bayesian priors
Externally visible characteristics
Spatial interpolation

ABSTRACT

DNA-based prediction of externally visible characteristics has become an established approach in forensic genetics, with the aim of tracing individuals who are potentially unknown to the investigating authorities but without using this prediction as evidence in court. While a number of prediction models have been proposed, use of prior probabilities in those models has largely been absent. Here, we aim at compiling information on the spatial distribution of eye and hair coloration in order to use this as prior knowledge to improve prediction accuracy. To this end, we conducted a detailed literature review and created maps showing the eye and hair pigmentation prevalence both by countries with available information and by interpolation in order to obtain prior estimates for populations without available data. Furthermore, we assessed the association between these two traits in a very large data set. A strong limitation was the quite low amount of available data, especially outside Europe. We hope that our results will facilitate the improvement of already existing and of novel prediction methods for pigmentation traits and induce further studies on the spatial distribution of these traits.

## 1. Introduction

Prediction of externally visible characteristics (EVC's) based on genetic data, often referred to as Forensic DNA Phenotyping (FDP), has become a major focus in forensic genetic research in the past years. The prime motivation behind this approach is to narrow down the group of potential trace donors in cases where standard genetic fingerprinting, such as short-tandem repeat (STR) profiling, could not provide any matching information with a priori known profiles [1,2]. FDP may thereby help focusing police investigations on a (limited) group of suspects, although the legal and ethical framework for such approaches is currently subject to intensive debate (see, for example, [3–7]).

Pigmentation traits, including eye, hair and skin coloration, have been subject of scientific studies for more than 130 years [8,9] and a primary focus of FDP in the past decade. Recent genome-wide association studies have identified numerous genetic variants that contribute to pigmentation [10–18] while others have subsequently used them to predict these traits [19–22]. Prominent and widely used approaches include IrisPlex [23,24] for eye color prediction and its extensions HIrisPlex [25,26] for combined eye and hair color prediction and HIrisPlex-S [27] for predicting the three pigmentation traits eye, hair and skin color. These models are based on multinomial logistic regression models where probabilities for each color category are presented for each trait. IrisPlex uses a total of 6 SNPs for eye color prediction while HIrisPlex includes a set of 24 SNPs and the lastest extension of HIrisPlex-S includes the already 24 established SNPs and additionally 17 related with the skin color prediction. Another established tool for eye color is Snipper [28] which is a Bayesian classifier and uses likelihood ratios to present the outcome of the prediction. All possible likelihoods are calculated and sorted in descending order and the final prediction is the ratio of the two largest likelihoods. Some alternative models are the ones developed by Söchtig et al. [29] and Maroñas et al. [30] for hair and skin color, respectively. These approaches are using iterative naïve Bayesian classification for pairwise phenotype differentiation. Lastly, classification tree approaches were also used for prediction, such as the model by Allwood et al. [31] for eye color prediction and 7-Plex and 8- Plex [32] for both eye and skin color.

In contrast to eye and hair color, which can often be predicted with high accuracy from relatively small sets of DNA markers, skin color prediction is frequently less powerful. Besides the few established tools for skin color prediction mentioned above, numerous studies in the past have already tried to identify genetic variation that underlies human skin coloration which can be, among other causes, a result of the exposure of UV radiation or natural selection [33–38]. Other studies indicated markers that are informative about skin pigmentation,

* Corresponding author at: Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany.
E-mail address: michael.nothnagel@uni-koeln.de (M. Nothnagel).

https://doi.org/10.1016/j.fsigen.2019.01.001
Received 13 June 2018; Received in revised form 11 December 2018; Accepted 1 January 2019
Available online 02 January 2019

Table 1
Eye color prevalence across Europe and Central Asia.

| Country | 'Blue' | 'Intermediate' | 'Brown' | References |
|---|---|---|---|---|
| Armenia | 3.05 (0.84-7.63) | 16.78 (10.83-24.31) | 80.15 (72.29-86.61) | [19] |
| Azerbaijan | 6.34 (1.76-15.47) | 22.21 (12.72-34.46) | 71.42 (58.65-82.11) | [19] |
| Denmark | 59.60 (56.67-63.22) | 23.70 (21.02-26.73) | 16.10 (13.87-18.83) | [43,50,53] |
| | 66.20 (57.89-73.84) | 27.60 (20.49-35.61) | 6.20 (2.87-11.45) | |
| | 69.50 (62.12-76.28) | 23.60 (17.47-30.58) | 6.90 (3.61-11.74) | |
| | 64.90 (57.15-72.07) | 25.00 (18.65-32.25) | 10.10 (6.00-15.70) | |
| | 58.50 (50.87-65.89) | 27.30 (20.84-34.48) | 14.20 (9.41-20.25) | |
| | 70.15 (65.29-74.70) | 11.78 (8.72-15.44) | 18.06 (14.33- 22.29) | |
| average | 64.84 | 20.45 | 14.50 | |
| Great Britain | 37.59 (29.34-46.40) | 18.79 (12.55-26.48) | 43.60 (35.03-52.47) | [51,53,54] |
| | 44.7 (42.38-47.13) | 29.9 (27.73-32.11) | 25.40 (23.32-27.49) | |
| | 46.1 (44.23-47.92) | 27.7 (26.03-29.35) | 26.30 (24.65-27.92) | |
| average | 42.80 | 25.46 | 31.77 | [51,53,54,72] |
| France | 22.00 | 44.00 | 34.00 | [65] |
| Georgia | 7.51(3.66-13.40) | 18.79 (12.55-26.48) | 73.68 (65.35-80.94) | [19] |
| Germany | 39.6 (39.58-39.65) | 33.2 (33.17-33.24) | 27.2 (27.15-27.22) | [9] |
| Iceland | 75.15 (74.93-78.03) | 12.95 (11.99-14.47) | 10.1 (9.22-11.45) | [12] |
| | 73.90 (74.05-77.35) | 15.35 (14.35-17.16) | 8.35 (7.52-9.69) | |
| average | 74.52 | 14.15 | 9.22 | |
| Kazakhstan | 3.33 (0.41-11.53) | 11.65 (4.82-22.57) | 85.00 (73.43-92.90) | [19] |
| Netherlands | 60.90 (61.97-67.59) | 11.40 (10.27-14.14) | 21.70 (20.65-25.63) | [12] |
| Poland | 52.50 (49.33-55.55) | 12.50 (10.49-14.64) | 35.10 (32.17-38.12) | [52] |
| Slovenia | 44.70 (35.05-54.78) | 25.70 (17.68-35.17) | 29.60 (21.02-39.22) | [55] |
| Tajikistan | 6.83 (3.00-13.03) | 7.67 (3.58-14.10) | 85.47 (77.76-91.30) | [19] |
| Ukraine (Crimea) | 25.00 (16.55-35.11) | 24.99 (16.55-35.11) | 50.00 (39.39-60.61) | [19] |
| Uzbekistan | 3.44 (0.95-8.59) | 6.02 (2.46-12.04) | 90.51 (83.66-95.17) | [19] |

Point estimates and, in parentheses, 95% confidence intervals for the trait prevalence are given in percentages. If more than one study was available per country, the average across those studies is given.

representing various sets of selected SNPs that were used later on for skin color prediction [10,32,39–44].

Bayesian approaches form a major, powerful and versatile class of statistical prediction models. These models combine probability density estimates of the groups to be distinguished with prior probabilities for the occurrence of subjects/objects from these groups, eventually obtaining posterior probabilities that are used to classify a given subject or object [45]. This use of priors represents a major strength of Bayesian models since it allows incorporating prior knowledge in the prediction, which may in turn improve prediction accuracy. Previous studies have already established some DNA based prediction models built upon naïve Bayesian likelihood classification or Bayesian networks [28–30,46–48].

By using priors in the data analysis, all prior information that is available for an unknown parameter can be expressed before any data-based evidence is considered. Priors can be obtained from past or external information, according to similar experiments, or may also express some subjective belief on the topic. In the context of FDP, priors may be obtained from trait prevalence in the general population, in a specific spatial, ethnic, social or religious subgroup, or in another relevant reference group. Inference of the biogeographic ancestry of an individual in particular may help identifying appropriate prior probabilities if prevalence values for the respective groups are available. However, currently existing approaches to FDP barely use prior knowledge of the biogeographic distribution of traits so far [3,23–26,28,32] and the potential advantage of using priors for appearance prediction has yet to be demonstrated. One example for the use of prior information is Snipper's so-called LOCPRIOR information, i.e. a numerical code that denotes shared characteristics between individuals of population groups and frequencies correlated among populations. The LOCPRIOR information indicates and provides information on specific traits within a population, such as linguistic, geographical, phenotypic ones etc [49]. In this case, sampling locations were used as prior information to assist the clustering between the different categories. Among other causes, the lack of prior implementation may be due to a simple lack of reliable data. While, on the other hand, several internet web sites and press articles present colorful maps on pigmentation traits and give impressive numbers for trait prevalence, their virtually always fail to name their data sources and used methods, rendering such information highly questionable.

In the present study, we aimed for compiling reliable prevalence data for two pigmentation traits, namely categorical eye and hair color, in European populations and beyond. To this end, we conducted a country-specific literature review in order to estimate the geographic prevalence distribution and to obtain reasonable prior probabilities that could be used for improving the prediction accuracy of pigmentation traits. Somewhat surprisingly, we found the available amount of reliable scientific data to be quite limited, despite an ongoing interest in this topic for over a century. We therefore report on only a limited set of mostly European countries for which we were able to compile trustworthy population-representative data concerning eye and hair color distribution. Furthermore, we performed spatial interpolation based on these data. Finally, we report on the extent of correlation between eye and hair color based on the largest study to date on this topic.

## 2. Materials and methods

### 2.1. Pigmentation color categories

In this study, we focused on categorical color categories for eye and hair color, motivated by the Bayesian approach of classifying subjects or objects into one of a set of distinct groups and also by the wider availability of such data compared to only recently introduced metric measurements of pigmentation. More specific, we used the categories 'blue', 'brown' and 'intermediate' for eye color description, as a simplification to the categories established in study of Sulem [12], namely blue/gray, green and brown/black. Categories 'blond', 'brown' and 'red' for hair color description were determined, following the studies of Lock-Andersen [43,50] and merging the brown and black coloration into one. A number of publications (both old and recent) deviated from this scheme. In order to keep the three-category classification and to ensure consistency across studies, we mapped 'hazel' eye color to 'brown', whereas 'green' and 'yellow' were mapped to 'intermediate'. The 'gray' color category was merged with 'blue' throughout.

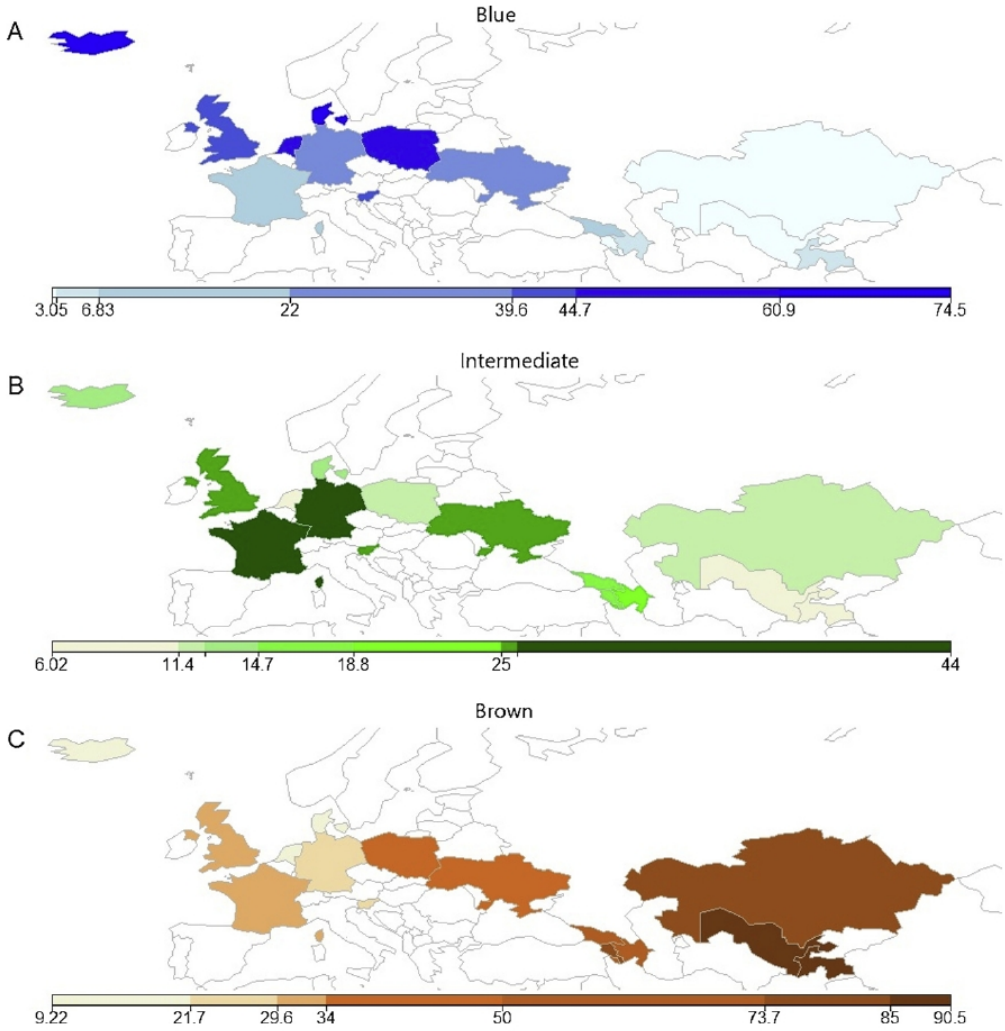


Fig. 1. Spatial distribution of categorical eye color prevalence across Europe and Central Asia. (A) 'Blue', (B) 'Intermediate', (C) 'Brown'. Numbers are given in percentages. Countries without available data are shown blank. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Correspondingly, 'black' and 'brown' hair color were amalgamated into a single category 'brown', whereas 'dark blond',' fair' and 'light brown' were merged with 'blond'. Shades such as 'auburn' or 'reddish' were already referred by a number of studies included, as one merged category with red hair. Separations such as 'light blue', 'dark blue' or 'grey' were aggregated into 'blue'.

### 2.2. Literature search and criteria for inclusion

We performed extensive literature searches (see Table S1 for a list of used terms) through PubMed (www.ncbi.nlm.nih.gov/pubmed) as well as through Google (www.google.com), and subsequently followed up references from publications obtained from these searches. Results fell into two distinct classes, namely comparatively recent human genetic studies and comparatively aged anthropological studies (see below). In order to assure reliable and population-representative data, we applied the following criteria to any source under consideration:

- Population-based sampling: We included only those studies where data sets were sampled in a population-based study or without respect to a specific pigmentation trait. In this way, we tried to avoid biases in the prevalence estimation that, for example, may occur with oversampling of individuals belonging to a certain category in retrospective association studies.
- Sample size: In order to protect against sporadic reports and strong chance deviations of the prevalence estimates from the population mean, we required each included study to have a minimum size. In particular, we chose a minimum size of 60 for eye color and of 130 for hair color. Uncertainty of the derived prevalence estimates was assessed by 95% confidence intervals. Studies whose total sample size failed the thresholds stated above were excluded.
- Plausibility: Studies had to be consistent in their presentation to be considered trustworthy. We encountered numerous studies where pigmentation category proportions did not sum up to unity (with totals both substantially below and above 100%) or inconsistent naming of categories. We therefore required studies to actually yield pigmentation category totals of 100% (or close) for consistently named categories in order to be included; otherwise they were

Table 2
Hair color prevalence across Europe.

| Country | 'Brown' | 'Blond' | 'Red' | References |
|---|---|---|---|---|
| Denmark | 29.00 (26.07-32.14) | 67.30 (64.07-70.33) | 3.60 (2.47-5.03) | [43,50,53] |
| | 11.10 (6.44-17.3) | 83.40 (76.38-89.10) | 5.50 (2.41-10.58) | |
| | 12.60 (8.14-18.61) | 79.30 (73.00-85.49) | 7.50 (4.06-12.51) | |
| | 9.50 (5.68-15.36) | 76.80 (71.59-84.66) | 11.30 (7.12-17.50) | |
| | 19.80 (14.34-26.70) | 76.10 (69.59-82.63) | 3.40 (1.27-7.31) | |
| | 49.73 (44.61-54.87) | 45.81 (40.73-50.95) | 4.45 (2.61-7.03) | |
| average | 30.66 | 64.00 | 4.99 | |
| Great Britain | 64.65 (55.91-72.75) | 20.30 (13.83-28.14) | 15.03 (9.43-22.26) | [51,53,54] |
| | 38.3 (36.10-40.75) | 56.40 (54.04-58.78) | 5.20 (4.17-6.33) | |
| | 47.5 (45.64-49.34) | 47.40 (45.54-49.24) | 5.10 (4.34-5.99) | |
| average | 50.15 | 41.36 | 8.44 | |
| Estonia | 43.00 | 56.00 | 1.00 | [65] |
| France | 84.00 | 12.00 | 4.00 | [65] |
| Germany | 31.40 (31.34-31.41) | 68.40 (68.36-68.43) | 0.20 (0.23-0.23) | [9] |
| Iceland | 26.20 (25.12-28.37) | 64.70 (64.28-67.75) | 7.10 (6.33-8.25) | [12] |
| | 26.00 (24.79-28.18) | 65.55 (64.87-68.48) | 6.75 (5.92-7.87) | |
| Netherlands | 25.90 (23.44-28.45) | 71.45 (68.84-74.00) | 2.60 (1.81-3.70) | [12] |

Point estimates and, in parentheses, 95% confidence intervals for the trait prevalence are given in percentage. If more than one study was available per country, the average across those studies is given.

excluded.

#### 2.2.1. Recent human genetic studies

A number of human genetic studies have been published in recent years. Perhaps not surprising given the higher levels of variety in Europe compared to other continents, most of these publications used European samples or those of European descent. These studies include genome-wide association studies or smaller studies on eye color [12,19,43,50–55] and on hair color [12,43,50,51,53,54,56]. Notably, some eye color data outside Europe were available for Central Asia for populations along the "Silk Road", namely from Armenia, Azerbaijan, Tajikistan, Kazakhstan, Georgia, Ukraine (Crimea) and Uzbekistan [19] (Supplementary Table S3).

Despite this apparent wealth of studies, at least on Europe, we had to exclude a number of them [57–64] for the potentially biased collection of their data, for comprising too small of a sample size, for lack of information on the data sources or for being not population-representative. In total, twelve studies met our quality criteria for population representativeness and were used for the subsequent analysis.

#### 2.2.2. Anthropological studies

Externally visible pigmentation traits have been subject to anthropological studies for over a century. It should be noted, however, that the interest was not in eye and hair color, or cranial and morphological measurements and other features in this respect, per se, but as means to define postulated basic human types, or "races", which would then be used to explain human appearance as a potential mixture of such types to varying degrees. While this originally scientific working hypothesis has long been dropped, and rightly so, and despite later or even contemporary distortions and misuses of this concept, best known from the race ideologies of the 19$^{th}$ and 20$^{th}$ century, those early publications may potentially inform on the prevalence of pigmentation traits in a number of different populations. We found a few studies that promised to serve this purpose, namely studies by Virchow [9], Galton [8] and Coon [65]. The impressive (and apparently largely forgotten) anthropological survey by the famous German physician Rudolf Virchow from the second half of the 19$^{th}$ century [9] represents by far the largest study on pigmentation traits ever conducted. An enormous sample of 6,758,827 children aged between 6 and 14 years old from different schools across all German provinces was collected, with schoolteachers assessing their pigmentation. This study also gave separate information on the German Jewish population, likely to be representative of the Ashkenazim, and provided evidence for less than one third of Germans being fair skinned with blond hair and blue eyes, for significant numbers of Prussians showing dark pigmentation and for a large proportion of German Jews being blond. Notably, Virchow arrived at the conclusion that there were no specific patterns in the pigmentation of any "race" under study. Within the framework of our analyses, we considered Virchow's data as being representative for today's Germany pigmentation distribution. However, some caution is required regarding the representativeness of the hair color prevalence values given the young age of the children in the survey and age-related hair color darkening in adulthood [1,66–71], rendering the estimate for blond hair an upper limit. Another, apparently useful, source was a book by Carleton S. Coon on "The Races of Europe" [65] and their physical anthropology. This book focused on the origin and the admixtures of purported human types, such as Upper Paleolithic, Caucasoid and Mediterraneans, and also aimed at compiling and synthesizing information from a vast number of different sources on features such as eye, hair and skin color, beard rufosity in males, cranial and nasal indexes and height, and more, covering an area that stretched from Northern Europe through Middle East, Central and West Europe, Mediterranean countries to North Africa and Western Asia. While at first apparently representing a comprehensive and reliable source, the variety of cited literature led to inconsistent, incomplete and frequently missing data. We finally considered only information on eye and hair color from France and on hair color from Estonia trustworthy and included it in our study. An early article by Francis Galton [8] provided some data on eye color in England, but they were not included due to unclear population representativeness.

#### 2.2.3. Web sites

We found a surprisingly large number of web sites, during the period from November through December 2017 that would apparently provide trait prevalence values for many populations and even interpolated colored geographic maps (see Table S2 for a list of such web sites). However, none of these sites provided information on the source of the presented data. We therefore excluded these web sites from further consideration based on their undocumented and therefore highly questionable data basis. The substantial number of web sites may reflect a strong interest of the general public in these traits, which thereby would be promising means to attract traffic to these sites.

### 2.3. Final data sets

Based on our selection criteria described above, we compiled two data sets from sources considered trustworthy, namely one for categorical eye color and one for categorical hair color. The eye color set

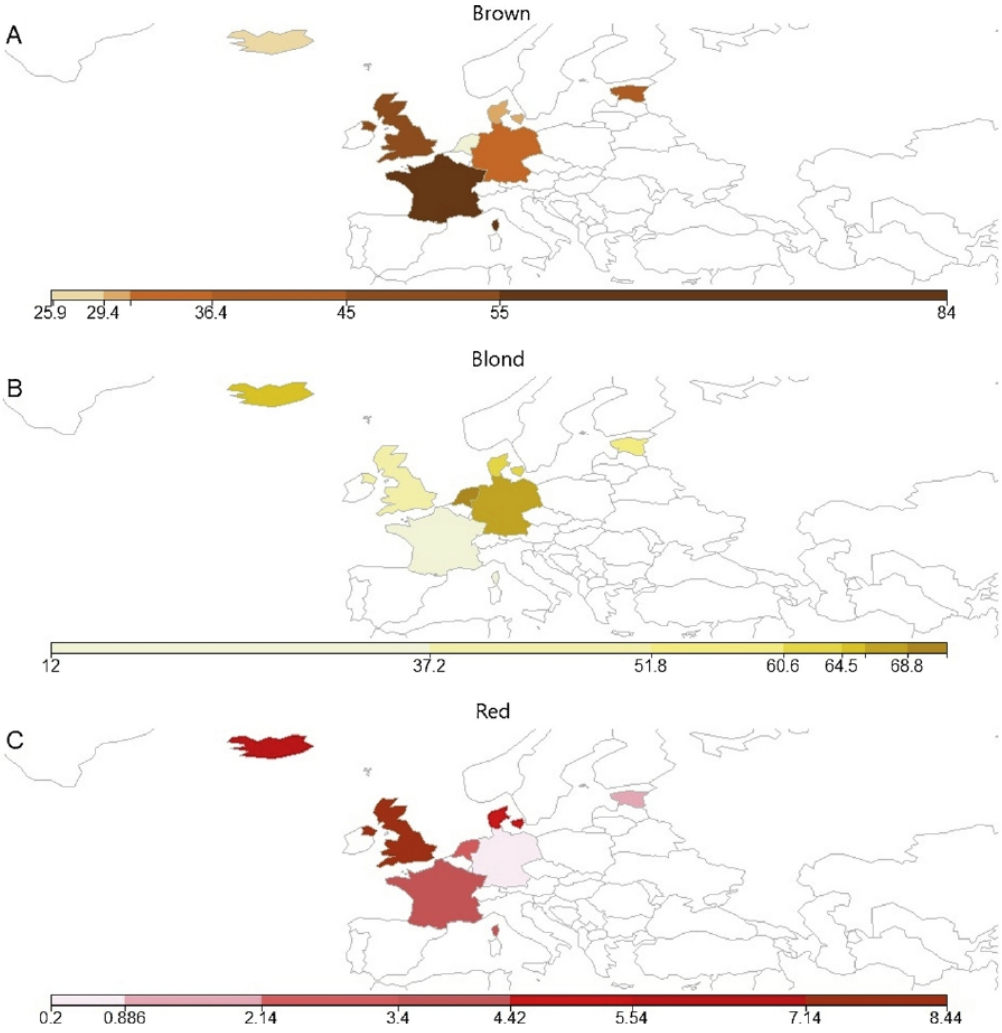


Fig. 2. Spatial distribution of categorical hair color prevalence across Europe. (A) 'Brown', (B) 'Blond', (C) 'Red'. Numbers are given in percentages. Countries without available data are shown blank. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

contained data from 16 European countries as well as from Central Asian countries along the "Silk Road" [9,12,19,43,50–55,65,72] (see Table S3 for the data summary). The hair color set was restricted to seven European countries, namely Denmark [43,50,53], Estonia [65], Great Britain [51,53,54], France [65], Germany [9] (see Table S4 for the overall data summary), Iceland [12] and The Netherlands [12]. If more than one study was available for a particular country, such as Denmark, we used the average prevalence for each category.

## 3. Statistical analysis

All analyses were performed in R version 3.4.3 [73]. Association between eye and hair color was assessed by Cramer's V on the 3 × 3 table of eye and hair color categories, while independence was tested by use of a chi-squared test where a P-value < 0.05 was deemed significant, as implemented in package lsr [74]. Uncertainty of proportion estimates for each category was assessed by binomial 95% confidence intervals, except for hair color in France and Estonia where information on the total sample size was lacking. Maps for visualization were created separately for eye and hair color and for each category, using packages vcd [75] and rworldmap [76]. Cut-offs were defined by use of quantiles. Spatial interpolation was performed to obtain trait prevalence estimates for geographic regions, or countries, for which no data were available. We employed inverse distance weighted (IDW) interpolation which estimates the value of an unmeasured location by using the values of the neighboring measured data points. Each of the values has a weight which is proportional to the inverse of the distance raised to a power function p. P can be any positive real number although in our case the default value of 2 was used. Since we do not have any directional influences in our data, the Shepard method was used which assumes equally influenced points in all directions. The general inverse distance weighted (IDW) interpolating function for finding an interpolated value $u$ at a given point $x$ based on the given points $u_i = u(x_i)$ for $i = 1,2, \ldots,N$ is of the form:

$$u(x) = \begin{cases} \frac{\sum_{i=1}^{N} w_i(x) u_i}{\sum_{i=1}^{N} w_i(x)} & : \; if\ d(x, x_i) \neq 0\ for\ all\ i \\ u_i & : \; if\ d(x, x_i) = 0\ for\ some\ i \end{cases}$$

where

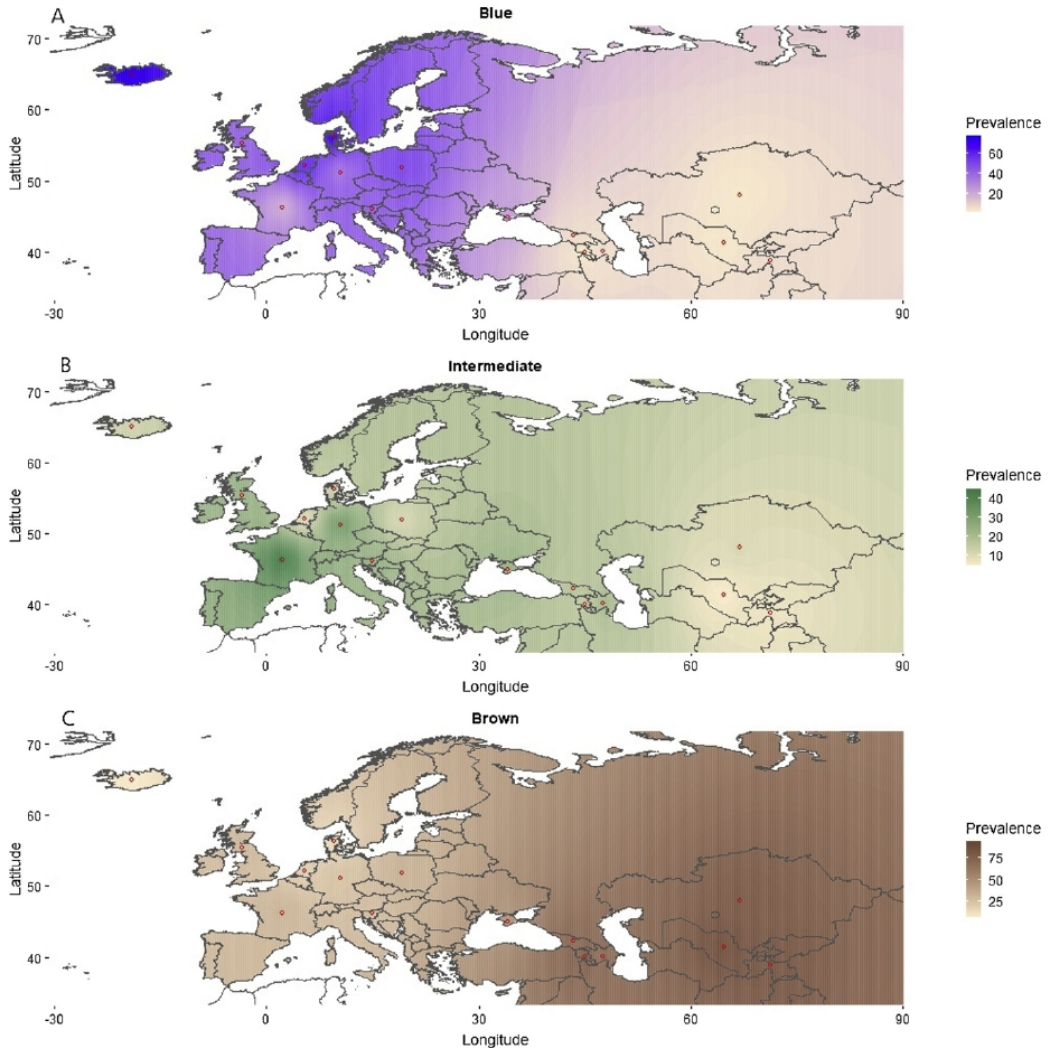


Fig. 3. Spatial interpolation of categorical eye color prevalence. (A) 'Blue', (B) 'Intermediate', (C) 'Brown'. Numbers are given in percentages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

$$w_i(x) = \frac{1}{d(x, x_i)^p}$$

is the inverse distance weighting function as defined by Shepard [77]. Variable $x$ represents an arbitrary interpolated point, $x_i$ a known interpolating point while $d$ is the distance between the known point $x_i$ and the unknown point $x$. $N$ denotes the number of the known data points. Due to the limited amount of data available for mapping, no specified number of neighbors was considered. The interpolation was performed by use of the gstat package [78,79], while ggplot2 [80] was used for creating graphs.

## 4. Results

### 4.1. Eye color prevalence

The prevalence estimates for 'blue', 'brown' and 'intermediate' eye color differed considerably between the countries for which reliable data were available, ranging from 3.05 to 90.51 (Table 1). The highest prevalence of any of the eye color traits was observed in Uzbekistan (> 90% for 'brown'), while Iceland followed closely with a prevalence of 74.52% for 'blue' eyes. 'Intermediate' color occurred less frequently compared to the other two categories and reached its maximum of 44% in France (Table 1). Prevalence of 'blue' eyes was smallest in Armenia (3.05%) among all countries with available data, whereas Iceland comprised the smallest prevalence (9.22%) of 'brown' eyes. In our study set, 'intermediate' eye color showed the lowest prevalence in Uzbekistan (6.02%). Overall, blue eye color dominated in Northern Europe, whereas brown eye color was highly prevalent in Central Asia (Fig. 1). Interestingly, France and Germany appeared to present the largest proportion of carriers of intermediately colored eyes among all countries with available data.

### 4.2. Hair color prevalence

Despite the limited amount of scientific data available, we were able to compile some trustworthy outcomes regarding the hair color prevalence for a small number of countries, mostly for the Northwestern part of Europe. Due to this fact, the results below present a preliminary

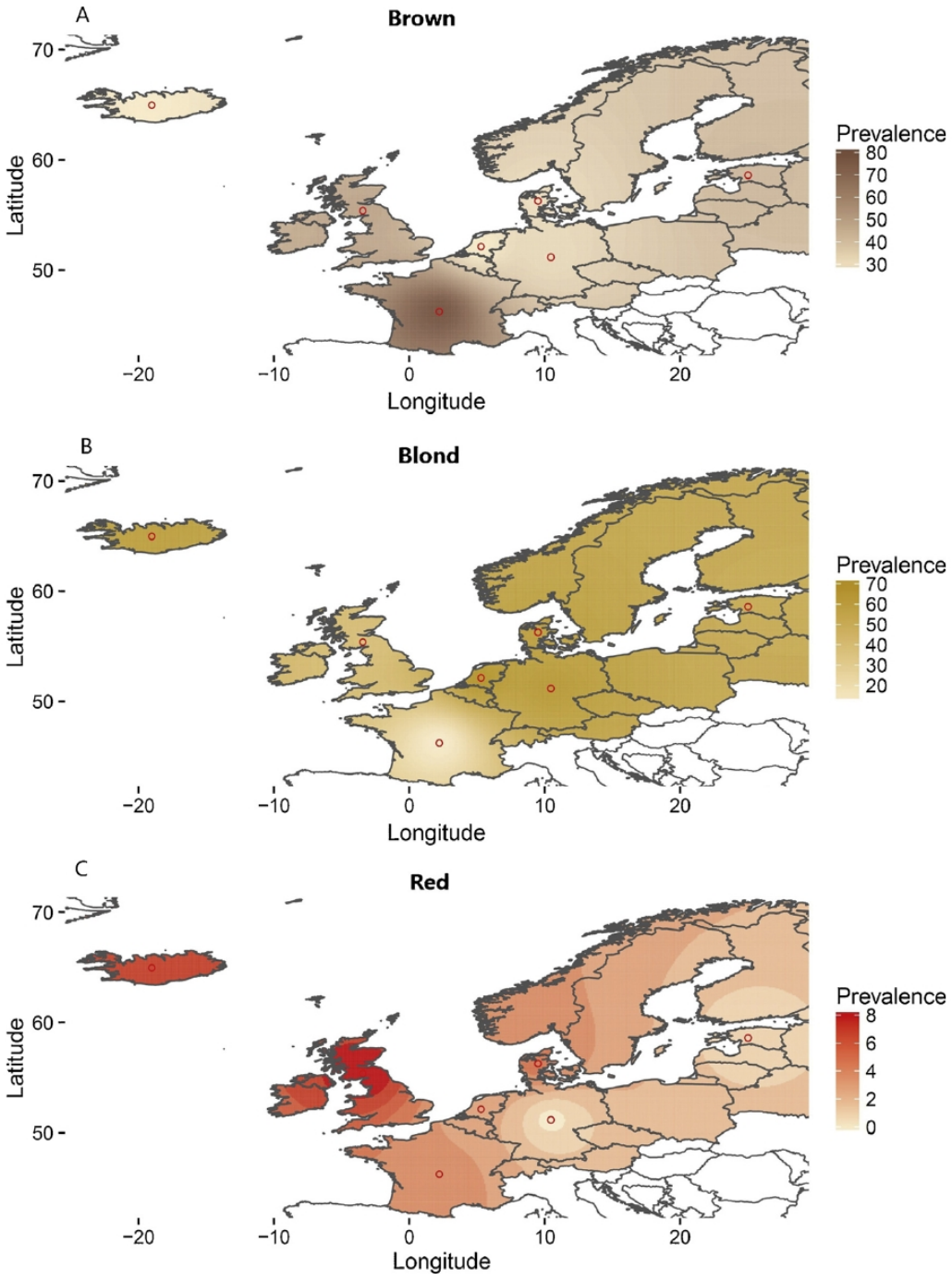


Fig. 4. Spatial interpolation of categorical hair color prevalence. (A) 'Brown', (B) 'Blond', (C) 'Red'. Numbers are given in percentages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

overview of the hair color hair frequency that should be considered with caution. Prevalence for all three color categories ('blond', 'brown' and 'red') ranged from 0.2 to 84% (Table 2). The highest prevalence was observed for 'brown' hair in France, simultaneously the smallest frequency of 'blond' hair (12%). The Netherlands showed the largest share of blond hair (71.45%) and the smallest prevalence of brown-haired people (25.9%). Red hair color prevalence was comparatively small throughout, being highest in Great Britain (8.44%) and quite low in Germany (0.2%; Table 2). Overall, we noticed that blond hair dominated in Germany and the Netherlands, while the red hair trait reached its most frequent occurrence in Iceland and Great Britain (Fig. 2). Great Britain also appeared to have, together with France, a

Table 3
Joint occurrence of eye and hair color in Germany.

| | | Hair color 'Blond' | 'Brown' | 'Red' | Total |
|---|---|---|---|---|---|
| Eye color | 'Blue' | 2,149,027 | 514,628 | 6657 | 2,670,312 |
| | 'Intermediate' | 1,582,339 | 650,468 | 5066 | 2,237,873 |
| | 'Brown' | 878,488 | 949,822 | 3864 | 1,832,174 |
| | Total | 4,609,854 | 2,114,918 | 15,587 | 6,740,359 |

Data were compiled from [9]. Numbers are given in counts.

Table 4
Hair color probability conditional on eye color category in Germany.

| | | Hair color 'Blond' | 'Brown' | 'Red' | Total |
|---|---|---|---|---|---|
| Eye color | 'Blue' | 80.4 | 19.2 | 0.24 | 99.84 |
| | 'Intermediate' | 70.7 | 29 | 0.22 | 99.92 |
| | 'Brown' | 47.9 | 51.8 | 0.21 | 99.91 |

Data were compiled from [9]. Numbers are given in percentage.

Table 5
Eye color probability conditional on hair color category in Germany.

| | | Eye color 'Blue' | 'Intermediate' | 'Brown' | Total |
|---|---|---|---|---|---|
| Hair color | 'Blond' | 46.6 | 34.3 | 19.05 | 99.95 |
| | 'Brown' | 24.3 | 30.7 | 44.9 | 99.9 |
| | 'Red' | 42.7 | 32.5 | 24.7 | 99.9 |

Data were compiled from [9]. Numbers are given in percentage.

high percentage of brown haired people. Denmark and Estonia had higher proportion of blond hair carriers compared to the other two hair color categories.

### 4.3. Prevalence interpolation

Interpolated maps, based on already collected information for the specific set of countries described above, may assist in the estimation of the eye and hair color prevalence for countries where no measurements were available. As one may notice, the blue eye color is predicted to dominate in Northern Europe compared to the Southeastern countries where the percentage resembles less than 10% (Fig. 3). The intermediate coloration was interpolated to be of higher prevalence in Western Europe and to spread almost uniformly across the Balkan countries. Carriers of brown eyes were interpolated with high frequency in the Eastern countries in contrast to Western Europe where we observed the smallest prevalence values. In Northern Europe, brown hair pigmentation occurs at minimum prevalence in contrast to blond hair which dominates there (Fig. 4). Red hair prevalence is predicted to be higher in Western compared to Eastern Europe.

### 4.4. Binomial confidence intervals

We also calculated, separately for each category, the 95% Binomial confidence intervals in order to assess the uncertainty of our prevalence estimates for each trait (Tables 1 and 2). It should be noted, however, that for countries with more than one study available, such as Denmark, confidence intervals were not always overlapping between studies.

### 4.5. Correlation between eye and hair color

Correlated traits may facilitate an improved prediction by borrowing strength from an already observed trait, although this is subject to ongoing investigation (not shown). We therefore investigated the degree of association between eye and hair color in the large German data set compiled by Virchow [9] (see Table S4 for an excerpt of the main results table). Based on the joint occurrence of eye and hair color categories in these data (Table 3), these pigmentation traits showed a modest association (Cramer's V = 0.20) that was nevertheless highly significantly different from zero ($p = 2.2 \times 10^{-16}$). We also estimated conditional probabilities for the presence of particular color categories in one trait given the presence of a category in the other trait, namely the probability of hair color given the eye color (Table 4) and eye color given the hair color (Table 5). Notably, as seen in Table 4, the 'blond' hair trait has a high conditional probability for 'blue' eye color (80.4%) but also for 'intermediate' eyes (70.7%). On the other hand, 'blond' and 'brown' hair colors are almost equally likely (47.9% vs 51.8%) given the presence of 'brown' eyes. The 'red' hair trait did not show a substantial conditional probability for any of the eye colors, which is not surprising giving the comparatively low prevalence of this trait in the population. More specifically, the conditional probability ranged between 0.21% and 0.24% but no significant difference occurred among the three eye categories. Given hair color, we observed a high conditional probability of 'blue' eyes with 'blond' hair, reaching the level of 46.6% (Table 5). This is followed by the 'intermediate' trait and lastly is the 'brown' eyes, which have only the 19% of our sample.

## 5. Discussion

In the present study, we aimed at assessing the spatial distribution of eye and hair pigmentation, separately for different countries. The basic motivation for this approach was the compilation of prior knowledge on this distribution for subsequent use in potentially enhancing the prediction accuracy of those two traits. To this end, we conducted a detailed literature review in order to collect information on the geographic prevalence distribution for eye and hair color. However, the lack of available data for many populations, especially outside of Europe, and the limited amount of reliable data were two significant obstacles to our analysis. Despite the fact that the distribution of human pigmentation has been a topic of major interest for decades, only few studies could be considered to provide trustworthy information concerning the origin of their data used and the representativeness of the sample. The majority of these studies were focused on the pigmentation of European populations, which immediately limited the scope of our research, evidencing the need for future studies with a worldwide research focus. In this way, it will be possible to estimate small prevalence values of exceptional pigmentations in populations which do not show large variability overall. Reports on blue-eyed people in the African and Chinese populations which, despite the low prevalence of their trait, may serve as an example; those occurrences should not be neglected from pigmentation distribution studies (see Supplementary Table S5).

Data on eye and hair pigmentation were provided by a number of studies, which focused on different topics such as genetic markers responsible for pigmentation [12,19,51–54], interdependence between pigmentation and incidence of diseases [50], prevalence of the distribution of hair and eye color in specific populations [43], anthropological studies on different populations [65] and prediction of eye color [55]. Here, we concentrated on the association between eye and hair coloration and also the pigmentation prevalence across different countries, mostly inside of Europe. It should be noted, however, that the obtained estimates are based on studies with usually quite limited sample sizes, as expressed by the considerable range of the corresponding confidence intervals which at times did not even overlap for the same country.

With respect to the eye and hair color in the data set on the German population by Virchow we noticed that the highest association was obtained between blond hair and blue eyes while brown hair was associated with brown eyes. The extremely high percentage of blond-haired individuals compared to the other two categories was remarkable in this data set but may be explicable not only by the fact that Germans predominantly feature light pigmentation, but also that the

sample set included only children aged between 6 and 14 years, whose pigmentation tends to be lighter than that of adults. Due to age-related hair color change, larger proportions of adults in future studies of Germany are likely to lead to smaller prevalence estimates for this pigmentation trait. Thus, hair color prevalence values in the adult German population will likely be slightly different from the ones reported in this study.

The generally lighter pigmentation in North Europeans compared to individuals from Eastern countries was to be expected. However, there were a few surprising cases where the frequency of a trait deviated from this expectation. This includes the highest percentage of intermediate eye color in France comprising almost the half of the population (44%) and also the extremely low percentage of red hair in Germany (0.2%). Furthermore, brown hair color was observed with higher frequency in France (84%) and also in the British population (50.15%). These discrepancies are explicable in terms of the size of the population data set, the age of the individuals included or the non-standardized way that these data were collected.

Throughout our study, we interpolated the pigmentation distribution for countries where no data were available. In the resulting maps we presented the estimations only for those countries that were considered trustworthy in our limited data set, especially for hair color. For eye color, some predictive errors are also likely to occur. In Italy, higher frequency of blue eyes compared to the other two categories was predicted. We could consider these values valid at least for the Northern part of Italy since data of neighboring countries such as France and Slovenia were available and included in our analyses, but not necessarily for Italy as a whole. It is therefore important to note that conclusions from our interpolation results should be drawn with caution. Thus, this study should be considered to be of preliminary nature and the information assembled here should be considered approximate for the definition of priors. More studies on pigmentation traits with a unified color category definition and larger sample sizes on a worldwide level will be required in order to arrive at a more accurate and comprehensive picture of pigmentation trait prevalence. This would allow to expand our results and provide prevalence estimates for continents and regions not yet covered. To be used in future prediction models, priors may be obtained by suitable averaging of prevalence values across countries, regions or even continents, depending on the degree of differentiation and the required accuracy. Given Europe's large internal variability, some applications may require prior values specific to Northern and Southern Europe or the continent as a whole. For this reason, we hope that studies on the topic that have been published in languages other than English or German and that unfortunately escaped our attention could be included in our data set in the future. This is likely to be of particular importance for the completion of the spatial picture of pigmentation trait prevalence for geographic regions outside Europe. Given the importance of data variation for the framework of this study, we surmise that the computational strategy followed here and the pigmentation maps presented will provide some assistance in different scientific fields and in future forensic application related to the EVC prediction.

## Funding

The authors received support from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 740580 within the framework of the Visible Attributes through Genomics (VISAGE) Project and Consortium. None of the funding organizations had any influence on the design, conduct or conclusions of the study.

## Conflict of interest

The authors declare that they have no competing interests.

## Authors' contributions

All authors read and approved the final manuscript.

## Acknowledgements

We thank Christian Andree, University of Kiel, Germany, and Georg Olms Verlagsbuchhandlung, Hildesheim, Germany, for the permission to reproduce Table "Die absoluten Gesammtergebnisse in den einzelnen Staaten des Deutschen Reiches" from Virchow's Collected Works [9] in electronic form as Table S4 in this publication. We also thank Sheila Ulivi for providing pigmentation data from Central Asian countries [19].

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.01.001.

## References

[1] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 7 (2013) 98–115.
[2] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.
[3] A. Caliebe, S. Walsh, F. Liu, M. Kayser, M. Krawczak, Likelihood ratio and posterior odds in forensic genetics: two sides of the same coin, Forensic Sci. Int. Genet. 28 (2017) 203–210.
[4] M. Kayser, Forensic DNA Phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.
[5] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (2009) 154–161.
[6] F. Staubach, Germany: note limitations of DNA legislation, Nature 545 (2017) 30.
[7] V. Toom, M. Wienroth, A. M'Charek, B. Prainsack, R. Williams, T. Duster, et al., Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively: reply to' Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes' by Manfred Kayser, Forensic Sci. Int. Genet. 22 (2016) e1–e4.
[8] F. Galton, Family likelness in eye-color, Proc. R. Soc. 40 (1886) 402–416.
[9] C. Andree (Editor and Processor): Rudolf Virchow. Sämtliche Werke. Bd. 45, Abt. III, Anthropologie, Ethnologie und Urgeschichte. Gesammtbericht über die von der deutschen anthropologischen Gesellschaft veranlassten Erhebungen über die Farbe der Haut, der Haare und der Augen der Schulkinder in Deutschland. Mit zusätzlichen Texten Virchows zur Forschungsgeschichte der Schulkindererhebungen, zur „Rassen"- und anthropologischen „Juden"-Frage. Erweiterte und verbesserte Ausgabe des Originals Berlin 1888. Hildesheim, Zürich, New York: Olms 2009.
[10] Liu Fan, M. Visser, L. Duffy, P.G. Hysi, C. Jacobs Leonie, O. Lao, et al., Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up, Forensic Sci. Int. Genet. 134 (2015) 823–835.
[11] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat. Genet. 40 (2008) 835–837.
[12] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, et al., Genetic determinants of hair, eye and skin pigmentation in Europeans, Nat. Genet. 39 (2007) 1443–1452.
[13] S.I. Candille, D.M. Absher, S. Beleza, M. Bauchet, B. McEvoy, N.A. Garrison, et al., Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four european populations, PLoS One (2012) 7.
[14] M.R. Gerstenblith, J. Shi, M.T. Landi, Genome-Wide Association Studies of Pigmentation and Skin Cancer: A Review and Meta-Analysis, HHS Author Manuscripts 23 (2010) 587–606.
[15] J. Han, P. Kraft, H. Nan, Q. Guo, C. Chen, A. Qureshi, et al., A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, PLoS Genet. (2008) 4.
[16] J. Oh, K. Zackowski, M. Chen, S. Newsome, S. Saidha, S.A. Smith, et al., Multiparametric MRI correlates of sensorimotor function in the spinal cord in multiple sclerosis, Mult. Scler. 19 (2013) 427–435.
[17] L. Rawofi, M. Edwards, S. Krithika, P. Le, D. Cha, Z. Yang, et al., Genome-wide association study of pigmentary traits (skin and iris color) in individuals of East Asian ancestry, PeerJ (2017).
[18] R.P. Stokowski, P.V.K. Pant, T. Dadd, A. Fereday, D.A. Hinds, C. Jarman, et al., A genomewide association study of skin pigmentation in a south asian population, Am. J. Hum. Genet. 81 (2007) 1119–1132.
[19] S. Ulivi, M. Mezzavilla, P. Gasparini, Genetics of eye colours in different rural populations on the Silk Road, Eur. J. Hum. Genet. 21 (2013) 1320–1323.
[20] Nicholas G. Crawford, Derek E. Kelly, Matthew E.B. Hansen, Marcia H. Beltrame,

Shaohua Fan, Shanna L. Bowman, et al., Loci associated with skin pigmentation identified in African populations, HHS Author Manuscripts (2017) 358.
[21] Liu Fan, Andreas Wollstein, Pirro G. Hysi, Georgina A. Ankra-Badu, Timothy D. Spector, Daniel Park, et al., Digital quantification of human eye color highlights genetic association of three new loci, PLoS Genet. (2010) 6.
[22] M. Kayser, L. Fan, C.J.W. Janssens, F. Rivadeneira, O. Lao, Kv Duijn, et al., Three genome-wide association studies and a linkage analysis identify HERC2 as a human Iris color gene, Am. J. Hum. Genet. 82 (2008) 411–423.
[23] S. Walsh, A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu, J. Seland, et al., DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 6 (2011) 330–340.
[24] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.
[25] S. Walsh, M. Kayser, A practical guide to the HIrisPlex system: simultaneous prediction of eye and hair color from DNA, Methods Mol. Biol. 1420 (2016) 213–231.
[26] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, et al., Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, Forensic Sci. Int. Genet. 9 (2014) 150–161.
[27] L. Chaitanya, K. Breslin, S. Zuñiga, L. Wirken, E. Pospiech, M. Kukla-Bartoszek, et al., The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: introduction and forensic developmental validation, Forensic Sci. Int. Genet. 35 (2018) 123–135.
[28] Y. Ruiz, C. Phillips, A. Gomez-Tato, J. Alvarez- Dios, M. Casares de Cal, R. Cruz, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2012) 28–40.
[29] J. Söchtig, C. Phillips, O. Maroñas, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, et al., Exploration of SNP variants affecting hair colour prediction in Europeans, Int. J. Legal Med. 129 (2015) 963–975.
[30] O. Maroñas, C. Phillips, J. Söchtig, A. Gomez-Tato, R. Cruz, J. Alvarez-Dios, et al., Development of a forensic skin colour predictive test, Forensic Sci. Int. Genet. 13 (2014) 34–44.
[31] Julia S. Allwood, S. Harbison, SNP model development for the prediction of eye colour in New Zealand, Forensic Sci. Int. Genet. 7 (2013) 444–452.
[32] K.L. Hart, S.L. Kimura, V. Mushailov, Z.M. Budimlija, M. Prinz, E. Wurmbach, Improved eye- and skin-color prediction based on 8 SNPs, Croat. Med. J. 54 (2013) 248–256.
[33] N.G. Jablonski, G. Chaplin, The evolution of human skin coloration, J. Hum. Evol. 39 (2000) 57–106.
[34] N.G. Jablonski, G. Chaplin, Epidermal pigmentation in the human lineage is an adaptation to ultraviolet radiation, J. Hum. Evol. 65 (2013) 671–675.
[35] G. Chaplin, Geographic distribution of environmental factors influencing human skin coloration, Am. J. Phys. Anthropol. 125 (2004) 292–302.
[36] P. Frost, Geographic distribution of human skin colour: a selective compromise between natural selection and sexual selection? Hum. Evol. 9 (1994) 141–153.
[37] J.H. Relethford, Hemispheric difference in human skin color, Am. J. Phys. Anthropol. 104 (1997) 449–457.
[38] A.R. Martin, M. Lin, J.M. Granka, J.W. Myrick, X. Liu, A. Sockell, et al., An Unexpectedly Complex Architecture for Skin Pigmentation in Africans, Cell 171 (2017) 1340–1353.
[39] M.D. Shriver, E.J. Parra, Dios Sonia, C. Bonilla, H.L. Norton, C. Jovel, et al., Skin pigmentation, biogeographical ancestry and admixture mapping, Hum. Genet. 112 (2002) 387–399.
[40] S. Walsh, L. Chaitanya, K. Breslin, C. Muralidharan, A. Bronikowska, E. Pospiech, et al., Global skin colour prediction from DNA, Hum. Genet. 136 (2017) 847–863.
[41] O. Spichenok, Z.M. Budmlija, A.A. Mitchell, A. Jenny, L. Kovacevic, D. Marjanovic, et al., Prediction of eye and skin color in diverse populations using seven SNPs, Forensic Sci. Int. Genet. 5 (2010) 472–478.
[42] B. McEvoy, S. Beleza, The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model, Hum. Mol. Genet. (2006) 15.
[43] J. Lock-Andersen, H.C. Wulf, N.D. Knudstorp, Interdependence of eye and hair colour, skin type and skin pigmentation in a Caucasian population, Acta Derm. Venereol. 78 (1998) 214–219.
[44] E.J. Parra, Human pigmentation variation: evolution, genetic basis, and implications for public health, Am. J. Phys. Anthropol. (Suppl 45) (2007) 85–105.
[45] G. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.
[46] E. Pospiech, J. Draus-Barini, T. Kupiec, A. Wojas-Pelc, W. Branicki, Prediction of eye color from genetic data using Bayesian approach, J. Forensic Sci. 57 (2012) 880–886.
[47] Kastelic Vanja, D. Katja, A single-nucleotide polymorphism (SNP) multiplex system: the association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model, Croat. Med. J. 53 (2012) 401–408.
[48] G.M. Dembinski, Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population, Forensic Sci. Int. Genet. 9 (2014) 111–117.
[49] Porras-Hurtado Liliana, Ruiz Yarimar, Santos Carla, Phillips Christopher, L.M.V. Carracedo Angel, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front. Genet. (2013) 4.
[50] J. Lock-Andersen, K.T. Drzewiecki, H.C. Wulf, Eye and hair colour, skin type and constitutive skin pigmentation as risk factors for basal cell carcinoma and cutaneous malignant melanoma. A Danish case-control study, Acta Derm Venereol. 79 (1999) 74–80.
[51] D.L. Duffy, G.W. Montgomery, W. Chen, Z.Z. Zhao, L. Le, M.R. James, et al., A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation, Am. J. Hum. Genet. 80 (2007) 241–252.
[52] E. Pospiech, J. Karlowska-Pik, B. Ziemkiewicz, M. Kukla, M. Skowron, A. Wojas-Pelc, et al., Further evidence for population specific differences in the effect of DNA markers and gender on eye colour prediction in forensics, Int. J. Legal Med. 130 (2016) 923–934.
[53] J. Mengel-From, T.H. Wong, N. Morling, J.L. Rees, I.J. Jackson, Genetic determinants of hair and eye colours in the Scottish and Danish populations, BMC Genet. 10 (2009) 88.
[54] D.L. Duffy, N.F. Box, W. Chen, J.S. Palmer, G.W. Montgomery, M.R. James, et al., Interactive effects of MC1R and OCA2 on melanoma risk phenotypes, Hum. Mol. Genet. 13 (2004) 447–461.
[55] V. Kastelic, E. Pospiech, J. Draus-Barini, W. Branicki, K. Drobnic, Prediction of eye color in the Slovenian population using the IrisPlex SNPs, Croat. Med. J. 54 (2013) 381–386.
[56] W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, S. Walsh, et al., Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443–454.
[57] M. Vaughn, R. van Oorschot, S. Baindur-Hudson, Hair color measurement and variation, Am. J. Phys. Anthropol. 137 (2008) 91–96.
[58] P.E. Lagouvardos, I. Tsamali, C. Papadopoulou, G. Polyzois, Tooth, skin, hair and eye colour interrelationships in Greek young adults, Odontology 101 (2013) 75–83.
[59] M. Steggerda, Change in hair colour with age, J. Hered. 32 (1941) 402–404.
[60] C. Burt, The relation between eye-colour and defective colour-vision, Eugen. Rev. 37 (1946) 149–156.
[61] A.R. Wielgus, T. Sarna, Melanin in human irides of different color and age of donors, Pigment Cell Res. 18 (2005) 454–464.
[62] A. Taylor, Eye colour in the Tristan Da Cuncha, Hum. Biol. Baltimore. 39 (1967) 316.
[63] A.G. Froelich, W.R. Stephenson, Does eye color depend on gender? It might depend on who or how you ask, J. Stat. Educ. (2013) 21.
[64] W. Branicki, U. Brudnik, A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype, Ann. Hum. Genet. 73 (2009) 160–170.
[65] C. Coon, The Races of Europe, Macmillan, 1939.
[66] M. Kukla-Bartoszek, E. Pośpiech, M. Spólnicka, J. Karłowska-Pik, D. Strapagiel, E. Ządzińska, et al., Investigating the impact of age-depended hair colour darkening during childhood on DNA-based hair colour prediction with the HIrisPlex system, Forensic Sci. Int. Genet. 36 (2018) 26–33.
[67] S. Aneta, Ż Elzbieta, R. Iwona, Effects of psychological stress on skin and hair pigmentation in Polish adolescents, Anthropol. Rev. 75 (2012) 1–17.
[68] S. Commo, K. Wakamatsu, I. Lozano, S. Panhard, G. Loussouarn, B.A. Bernard, et al., Age-dependent changes in eumelanin composition in hairs of various origins, Int. J. Cosmet. Sci. 34 (2012) 102–107.
[69] H. Bogaty, Differences between adult and children's hair, J. Soc. Cosm. Chem. 20 (1969) 159–171.
[70] M. Trotter, H.L. Dawson, The Hair of French Canadians, Am. J. Phys. Anthropol. 18 (2005) 443–456.
[71] M. Trotter, The form, size, and color of head hair in American whites, Am. J. Phys. Anthropol. 14 (1930) 433–445.
[72] F. Galton, Family likeness in Eye colour, R. Soc. 40 (1886) 402–416.
[73] R develompent core team, The R Project for Statistical Computing, (2018).
[74] D. Navarro, lsr: Companion to "Learning Statistics with R", (2015).
[75] D. Meyer, vcd: Visualizing Categorical Data, (2017).
[76] A. South, rworldmap: A New R package for Mapping Global Data, R J. 3 (2011) 35–43.
[77] S. Donald, A two-dimensional interpolation function for irregularly-spaced data, ACM' 68 Proceedings of the 1968 23rd ACM National Conference, (1968), pp. 517–524.
[78] E.J. Pebesma, Multivariable geostatistics in S: the gstat package, Comput. Geosci. (2004) 683–691.
[79] B. Gräler, E.J. Pebesma, G. Heuvelink, Spatio-Temporal interpolation using gstat, R J. 8 (2016) 204–218.
[80] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.

## 11.2 Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits

This text was published as an article in Forensic Science International Genetics in 2021, 50, 102412, https://doi.org/10.1016/j.fsigen.2020.102412, 

Forensic Science International: Genetics 50 (2021) 102412

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

# Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits

Maria-Alexandra Katsara [a], Wojciech Branicki [b,c], Ewelina Pośpiech [b], Pirro Hysi [d], Susan Walsh [e], Manfred Kayser [f], Michael Nothnagel [a,g,*], on behalf of the VISAGE Consortium

[a] Cologne Center for Genomics, University of Cologne, Cologne, Germany
[b] Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[c] Central Forensic Laboratory of the Police, Warsaw, Poland
[d] Department of Twin Research & Genetic Epidemiology, St Thomas Hospital, Campus, Kings College London (KCL), London, UK
[e] Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, USA
[f] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands
[g] University Hospital Cologne, Cologne, Germany

ARTICLE INFO

*Keywords:*
Appearances
Genetic prediction
Impact of priors
Predictive DNA analysis
Forensic DNA phenotyping
Externally visible characteristics

ABSTRACT

The prediction of appearance traits by use of solely genetic information has become an established approach and a number of statistical prediction models have already been developed for this purpose. However, given limited knowledge on appearance genetics, currently available models are incomplete and do not include all causal genetic variants as predictors. Therefore such prediction models may benefit from the inclusion of additional information that acts as a proxy for this unknown genetic background. Use of priors, possibly informed by trait category prevalence values in biogeographic ancestry groups, in a Bayesian framework may thus improve the prediction accuracy of previously predicted externally visible characteristics, but has not been investigated as of yet. In this study, we assessed the impact of using trait prevalence-informed priors on the prediction performance in Bayesian models for eye, hair and skin color as well as hair structure and freckles in comparison to the respective prior-free models. Those prior-free models were either similarly defined either very close to the already established ones by using a reduced predictive marker set. However, these differences in the number of the predictive markers should not affect significantly our main outcomes. We observed that such priors often had a strong effect on the prediction performance, but to varying degrees between different traits and also different trait categories, with some categories barely showing an effect. While we found potential for improving the prediction accuracy of many of the appearance trait categories tested by using priors, our analyses also showed that misspecification of those prior values often severely diminished the accuracy compared to the respective prior-free approach. This emphasizes the importance of accurate specification of prevalence-informed priors in Bayesian prediction modeling of appearance traits. However, the existing literature knowledge on spatial prevalence is sparse for most appearance traits, including those investigated here. Due to the limitations in appearance trait prevalence knowledge, our results render the use of trait prevalence-informed priors in DNA-based appearance trait prediction currently infeasible.

## 1. Introduction

Prediction of externally visible characteristics (EVCs) of an individual solely based on genetic information, also referred to as DNA phenotyping or forensic DNA phenotyping (FDP), has become a focus in human genetic research and applications, such as in forensics, ancient DNA analysis and other areas. In forensic cases where conventional DNA-profiling methods, typically based on short tandem repeat (STR) markers, fail to identify the crime scene sample donor, because the evidential DNA-profile does not match the DNA-profile of any of the case suspects or anybody in the criminal offender DNA database, FDP may provide significant leads for police investigations to find unknown

* Corresponding author at: Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany.
*E-mail address:* michael.nothnagel@uni-koeln.de (M. Nothnagel).

https://doi.org/10.1016/j.fsigen.2020.102412
Received 15 April 2020; Received in revised form 9 September 2020; Accepted 25 October 2020
Available online 4 November 2020

perpetrators [1–3]. In such cases, FDP can contribute significantly by narrowing down a potentially large number of putative sample donors to a smaller group of individuals that carry the FDP-derived EVC information on which the police can then focus with further investigation. Groups that do not carry such information can be left out from the police investigation. Thus far, for eye, hair and skin color various underlying genes have been identified, predictive DNA markers have been identified, DNA tests suitable for analyzing such genetic markers in forensic DNA samples and statistical prediction models have been developed [4–10], and some of these DNA test systems have been forensically validated [9,11,12]. For traits such as freckles and hair structure, some associated genetic markers and the first predictive models have already been published, respectively [13–16]; however, no forensically validated tool has been established so far. Prediction models for some other EVCs are currently under investigation [17–20].

Categorical prediction of eye, hair and skin color is often based on multinomial logistic regression (MLR) using established genetic marker panels. For instance, the IrisPlex test and model for eye color prediction consists of a set of 6 single-nucleotide polymorphisms (SNPs) [4,9,21]. Its extension to eye and hair color, the HIrisPlex test and model is based on 24 SNPs in total [11]. The latest extension is the HIrisPlex-S test and model, which consists of 41 SNPs and allows simultaneous prediction of eye, hair and skin color from a DNA sample [12]. All three prediction models are publicly available via https://hirisplex.erasmusmc.nl/. An alternative statistical tool for the prediction of eye, hair and skin color from genotype data is offered by Snipper [8,22,23], which uses pairwise likelihood ratios to present prediction outcomes, while other pigmentation prediction tool models were also developed (see [24] for a review). While some of these models show high prediction accuracies for some pigmentation categories, more research is currently under way in order to improve existing tools, either by including more SNP predictors after they have been identified in large-scale gene mapping studies, or by using alternative prediction methods.

Bayesian classification is a statistical approach that considers the data-independent probability of each category, or class, as well as the data-derived likelihood that a given subject or object belongs to a particular category, and bases the classification decision upon these probabilities. More specifically, the Bayesian approach combines a prior probability distribution on the different categories with the density probabilities obtained from the observed samples, yielding the posterior distribution used to predict category, or class, membership of an individual or object [25]. Prior probabilities for parameters may reflect previous evidence, but also purely subjective assessment or available information on these parameters from the past, before any evidence from the sample set at hand is considered. Incorporation of such prior knowledge in the data analysis may potentially increase the prediction accuracy, namely in situations where the prediction model does not include all causal genetic factors and where the environment contributes significantly to the trait variance via non-genetic factors. In both situations, trait prevalence-informed priors may then act as proxies for the yet unknown causal genetic factors and non-genetic factors in a population, group or region. In the framework of appearance DNA prediction, including FDP, inference of the biogeographic ancestry of an unknown DNA sample from which EVCs are to be predicted, together with the use of the trait class prevalence in such biogeographic ancestry group as prior in the EVC prediction model may improve the prediction accuracy. However, despite the already existing approaches for EVC prediction, the impact of trait prevalence priors on EVC prediction accuracies has not been investigated thus far.

For putting prior-based EVC prediction into practice within the concept of FDP, one would envision to first carry out forensic DNA ancestry testing on the unknown crime scene DNA samples and use the obtained ancestry outcome as guidance for allocating the appropriate trait class prevalence data for the EVC to be predicted, and finally use them as priors in EVC prediction. Based on the DNA-identified geographic region of ancestry of the tested DNA donor, allocated trait class prevalence data for different populations from such region would be averaged (or combined in another suitable way), in order to likely represent continental or sub-continental groups, and would then be used as priors for Bayesian EVC prediction on the same DNA sample previously used for ancestry testing. Alternatively, to avoid population averaging, DNA ancestry testing would need to be specific for a particular population, which not only requires the availability of trait prevalence data for such population but also the ability of forensic DNA ancestry to work on the population level.

Here, we assess the impact of incorporating prior knowledge on EVC trait prevalence in a Bayesian setting on improving the accuracy of DNA-based EVC prediction, but also potential pitfalls caused by misspecification of such prior probabilities. To this end, we consider EVCs such as eye, hair and skin color for which prior-free genetic prediction models have previously been established [9,11,12], but also traits such as hair structure and freckles for which the first prediction models were recently proposed without considering priors [13,15,16]. Given the sparsity or even lack of spatial or population-specific prevalence information available for each of these EVCs [24], we investigated the impact of prevalence-informed priors across a grid in the complete space of all possible values for each trait category, thereby emulating the (mis-) specification of the informative prior values. Prediction modelling was performed by applying previously proposed DNA predictors in datasets from different populations inside and outside of Europe. We report on standard prediction performance measures for each trait category separately and for all model measurements, and then compare prior-informed model-based prediction against prior-free models. Furthermore, we demonstrate the effect of priors on the overall prediction accuracy of the EVCs investigated.

## 2. Materials and methods

### 2.1. Data sets

For prediction modelling of eye color, hair color, skin color, hair structure and freckles we used various datasets, most of which were used previously for predicting these EVCs. For *eye, hair and skin color* we applied datasets that were part of the previously used data to establish the IrisPlex model for eye color, the HIrisPlex model for hair color, and the HIrisPlex-S model for skin color prediction, comprising of samples from different continental ancestries [9,11,12]. In particular, we used 1095 samples for eye, 1702 for hair and 1318 for skin color prediction (Table 1). For *hair structure*, we applied data from 2043 samples from different ancestries that were previously used as model testing dataset in the EUROFORGEN study on hair structure prediction [15]. Finally, for *freckles*, we used data from 1801 unrelated samples from the TwinsUK dataset, comprising European individuals from the United Kingdom [26]. For all traits, the available datasets were split into 80 % for model training and 20 % for model validation (Table 1).

As genetic markers in the prediction modelling, we used previously established DNA predictors for eye, hair, skin color, hair shape and freckles, respectively. In particular, for eye color prediction, we used the 6 SNPs from the previous IrisPlex eye color model [9]; for hair color prediction we used the 22 hair color informative SNPs from the previous HIrisPlex hair color model [11]; for skin color prediction, we used the 36

**Table 1**
EVC-specific data sets used for prediction model training and testing with and without the use of prevalence-informed priors.

| Appearance trait | Training set (80 %) | Test set (20 %) | References |
|---|---|---|---|
| **Eye color** | 876 | 219 | [9,11,12] |
| **Hair color** | 1361 | 341 | [9,11,12] |
| **Skin color** | 1054 | 264 | [9,11,12] |
| **Hair Structure** | 1634 | 409 | [15] |
| **Freckles** | 1440 | 361 | [26] |

skin color informative SNPs from the previous HIrisPlex-S skin color model [12]; for hair shape prediction, we used the 38 SNPs from the previous EUROFORGEN study on hair shape prediction [15]; and for freckles prediction, we used the 13 out of the 22 SNPs recently proposed for this purpose by Kukla-Bartoszek [13]. Not using the remaining 9 previously proposed freckles DNA predictors is explained by data availability and quality control issues (see below). Samples with incomplete genotype information per each EVC were excluded from our analysis.

### 2.2. Appearance trait categories

We considered the following trait categories:

- Eye color: Blue, Intermediate, Brown
- Hair color: Blond, Brown, Red, Black
- Skin color: Very Pale, Pale, Intermediate, Dark, Dark to Black
- Hair structure: Straight, Wavy, Curly
- Freckles: Freckled, Non-freckled

All traits were treated as categorical variables and were coded as '1', '2', '3' etc. up to the number of considered categories, which ranged between two for the presence or absence of freckles and five for skin color. In the course of our study, five-class problems turned out to be extremely computationally expensive and prohibitive for a comprehensive analysis. To overcome this problem, we reduced the 5-class category problem for skin color into two 4-class problems by either merging the first two categories very pale and pale or the last two dark or dark to black. Predictive DNA markers were considered under an allele-based model and, correspondingly, numerically coded as 0 for homozygosity of the major, i.e. more frequent, allele, 1 for heterozygosity and 2 for homozygosity of the minor, i.e. less frequent, allele. We did not consider interaction terms in the prediction models, as recently proposed for instance for freckles [13], in order to allow a consistent derivation of the posterior probabilities in the Bayesian approach across all EVCs. That means that for all EVCs, the models were defined considering the additive effects of the corresponding genetic markers.

### 2.3. Data cleaning

All data sets had undergone previously described quality control [9,11,12,15] and could be readily used in the prediction models, except for the TwinsUK data set for freckles prediction. For this reason, we applied standard quality control on the raw Twins UK data in order to be able to use them further in our analysis. For the freckles prediction we considered the markers recently proposed from Kukla-Bartoszek [13]; however, only 14 out of the previously reported 22 markers were available in the TwinsUK dataset we received for this study up on request from the Department of Twin Research, King's College London, of which 13 passed the quality-control and were thus used for freckles prediction modeling. More specific, we intended to remove markers that showed a strong deviation from Hardy-Weinberg equilibrium ($p < 0.001$), excessive heterozygosity (>0.001) [27], more than two alleles, an imputation info score of less than 0.8 or very low minor allele frequencies (MAF < 0.01). One of the markers did not pass this step of quality control, and was thus excluded from our analysis. Out of sample pairs with excessive identity-by-descent (IBD) allele sharing (>0.2), one randomly selected sample was removed in order to assure (approximate) independence. Finally, we performed a principal-components analysis (PCA) on the merged data set of TwinsUK and the complete dataset of the 1000 Genomes population data [28], comprising known ancestry, in order to identify and subsequently remove all samples with large-scale differences in ancestry. The latter were defined by the first two principal components, which were sufficient to cluster the individuals in population groups ($PC1 \geq 0.01$ and $PC2 \leq -0.02$). From this data set, we extracted those performed using PLINK v1.9 [29] and 'RStudio' v 3.4.4 [30].

### 2.4. Statistical analysis

#### 2.4.1. Prior-free trait prediction

For the prediction of eye, hair and skin color, we used standard multinomial logistic regression (MLR), as established by Liu et al. [4]. We also used MLR for the three-class problem of predicting hair structure [15], whereas standard binomial logistic regression (BLR) was used for predicting freckle presence or absence [13]. Individuals were predicted, or classified, as presenting with a specific trait category according to the highest posterior probability across all categories, with no minimal threshold imposed on this probability, although being explicitly equivalent to a minimum threshold of 1 by the number of trait categories. For all traits included in our study, each of the trait-specific data sets was randomly split into two independent subsets (Table 1), with 80 % being used for model training (training set) and 20 % for model prediction (test set).

#### 2.4.2. Prior-incorporated trait prediction

In the absence of detailed trait prevalence information on virtually all externally visible characteristics (EVCs) considered here for different populations or continental groups, we sought to assess the impact of priors on the prediction performance by exhaustively exploring the space of all possible tupels, i.e. an ordered list with respect to categories, of prior probability values. More specific, we performed Bayesian classification based on either MLR or BLR, again depending on the number of trait classes, by including the prior information in the calculation of the posterior probabilities. For a 3-class trait, the model was formed as follows [4]:

$$\ln\left(\frac{p_2}{p_1}\right) = \alpha_2 + \sum_{j=1}^{k} \beta_2(\pi_2)_j x_j$$

$$\ln\left(\frac{p_3}{p_1}\right) = \alpha_3 + \sum_{j=1}^{k} \beta_3(\pi_3)_j x_j$$

where the $p_i$ ($i = 1,2,3$) denote the probabilities of each category and $\alpha_i$, $\beta_i$ ($i = 2,3$) the respective regression coefficients, with the first category being used as reference, while $(\pi_1, \pi_2, \pi_3)_{\sum_{i=1}^{3}\pi_i=1}$ forms the tupels of prior values for the three categories and $k$ refers to the number of genetic markers included in the model, e.g. $k = 6$ for the IrisPlex model, whereas $j$ is an index referring to those genetic markers. Estimates for $\alpha_i, \beta_i$ were obtained by the MLR model from the respective training data sets (Table 1). Analysis was conducted in R version 3.4.3 [31] by using the nnet R package [32]. Following the standard Bayesian prediction framework, posterior probabilities were then obtained as the product between the data-dependent likelihood and the prior information:

$$\frac{\tau_2}{\tau_1} = \frac{\pi_2}{\pi_1} \times \frac{f_2(x)}{f_1(x)}$$

$$\frac{\tau_3}{\tau_1} = \frac{\pi_3}{\pi_1} \times \frac{f_3(x)}{f_1(x)}$$

where $\tau_i, \pi_i, f_i$ ($i = 1,2,3$) denote the posterior probabilities, the prior probabilities and the likelihoods for each of the three categories, respectively. From the above formulas, the posterior probabilities for each trait category were eventually obtained as:

$$\pi_2 = \frac{\exp\left(\ln\left(\frac{p_2}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right)}{1 + exp\left(\ln\left(\frac{p_2}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right) + exp\left(ln\left(\frac{p_3}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right)}$$

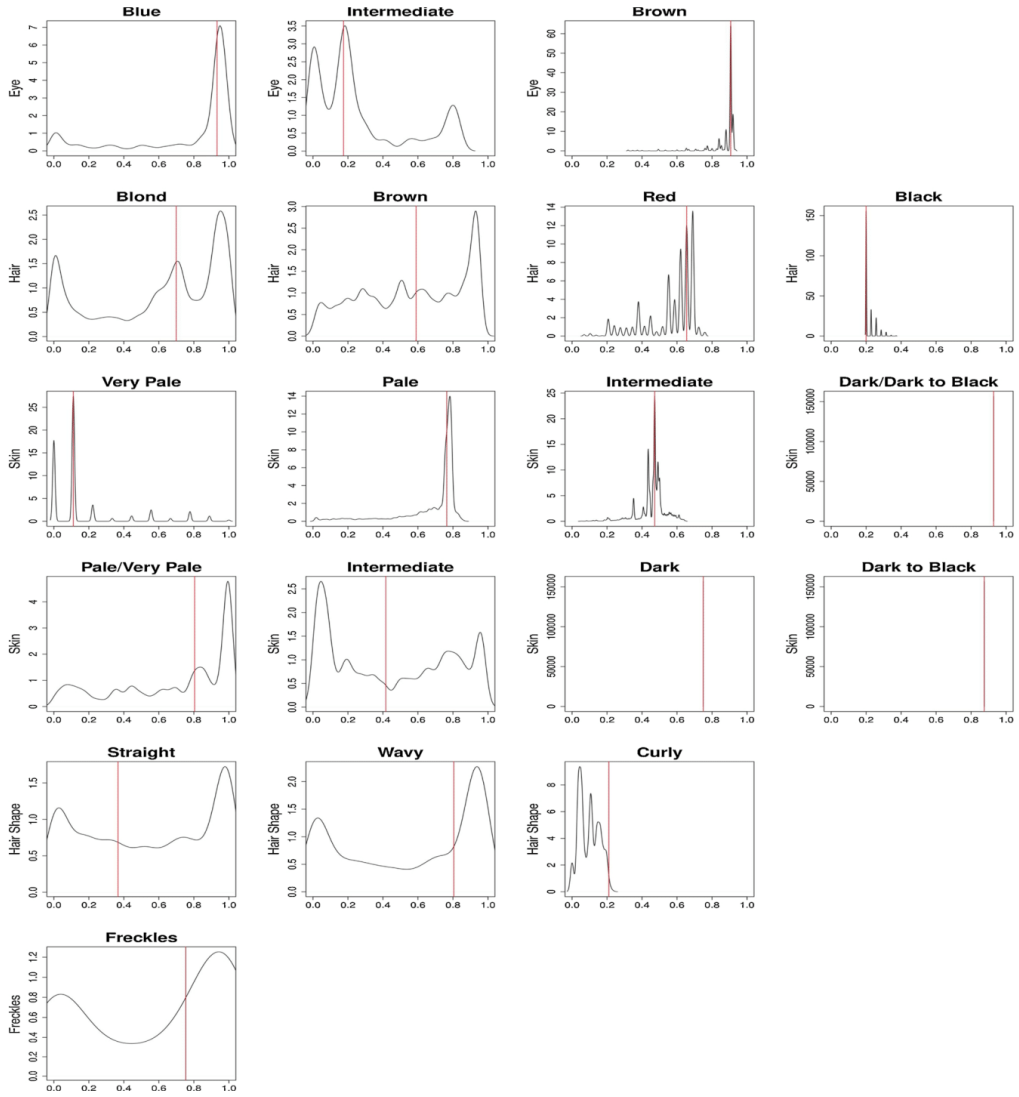


**Fig. 1. Impact of the choice of trait prevalence priors on sensitivity in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

$$\pi_3 = \frac{\exp\left(\ln\left(\frac{p_3}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right)}{1 + exp\left(\ln\left(\frac{p_3}{p_1}\right) + a_3 + \sum_{j=1}^{k}\beta_3(\pi_3)_j x_j\right) + exp\left(ln\left(\frac{p_3}{p_1}\right) + a_2 + \sum_{j=1}^{k}\beta_2(\pi_2)_j x_j\right)}$$

$$\pi_1 = 1 - \pi_2 - \pi_3$$

where $x_j$ denotes the number of minor (less frequent) alleles of the $j$th SNP and the terms $\alpha_i$ and $\beta_i$ $(i = 2, 3)$ are the model parameters. As before, indicator $j$ in the sum denotes the sum across all genetic markers. For simplicity, we did not consider interaction terms. This renders our approach only an approximation for the previously published freckles model. Models for the 2- and 4-class problems were defined in a similar fashion. A sample was classified into that category which yielded the maximum posterior probability, again without explicitly applying any minimal threshold.

With lacking trait prevalence information, we exhaustively explored the impact of priors by considering all possible tupels of prior probabilities in order to assess potential prediction improvement but also the risk caused by mis-specifying prior values. To this end, prior probabilities in turn assumed values from 0.01 to 0.99, with step size 0.01, while

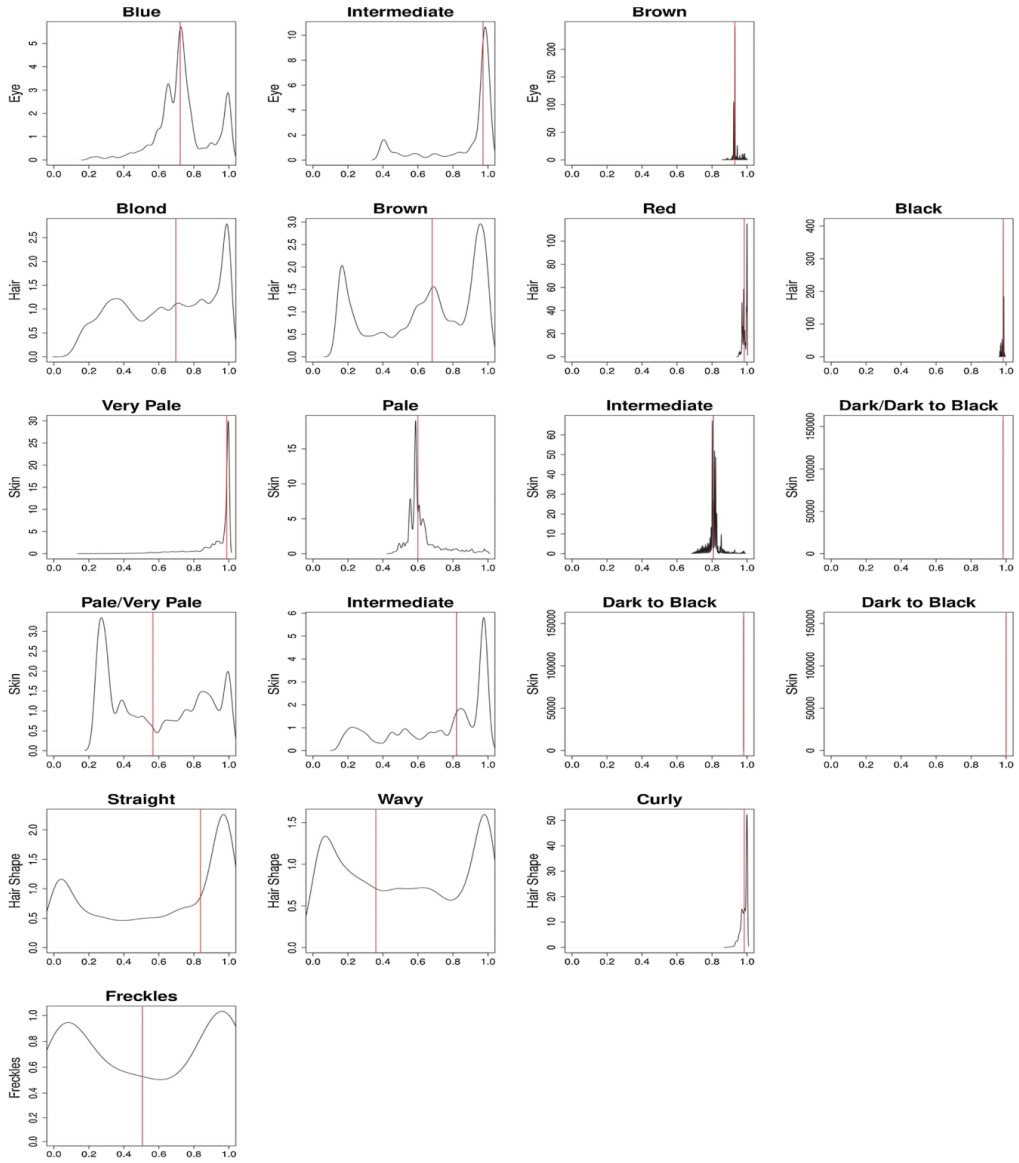


**Fig. 2. Impact of the choice of trait prevalence priors on specificity in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

requiring that those probabilities for all categories sum to unity. Note that prior-free prediction is equivalent to a Bayesian prediction model where the prior probabilities correspond to the relative trait category frequencies in the training set.

*2.4.3. Prediction performance assessment*

Prediction performance was evaluated in the respective test data sets (Table 1). We calculated commonly used measures of test accuracy, namely sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area-under-curve (AUC) and overall accuracy for all possible tupels of prior probabilities and subsequently summarized their distribution. In lay terms, sensitivity denotes the proportion of correctly predicted samples among all who manifest the trait category of interest (true-positive rate), whereas specificity denotes the proportion of correctly predicted samples among all that do not manifest the trait category of interest (true-negative rate). On the other hand, a PPV

refers to the proportion of correct classifications among all predictions of the trait category of interest, while an NPV refers to the proportion of correct classifications among all predictions other than the trait category of interest. The AUC denotes the area under the receiver-operating characteristic (ROC) curve which is obtained by varying the threshold used for the classification decision and can be interpreted as an indicator for the separability of classes when using the particular classification model. Except for freckles, which comprise only two categories, we performed multiclass ROC analysis, which carries out pairwise comparisons across all categories (one class vs. all other classes). For example, for eye color the following comparisons were conducted: "Blue vs. Non-Blue", "Intermediate vs. Non-Intermediate" and "Brown vs. Non-Brown". Finally, the overall model accuracy refers to the number of correct predictions divided by the total number of predictions made. As pointed out by Caliebe et al. [33], FDP does not operate in a diagnostic-test environment where (bio-)markers are being used to infer the presence of causal factors or conditions and where prediction is done in the opposite direction of causation. Instead, causal genetic markers, or proxies thereof, are used to predict the outcome along the causal direction. Thus, for FDP the most relevant performance measures are the predictive values (PPV and NPV). All analyses were performed in R v3.4.3 [30], using the packages nnet [32] and caret [34] for model building and performance assessment calculation, respectively, and package caTools [35] for multiclass AUC calculation. For visualization of our results, we used the package plot3D [36] and for better interpretability the standard kernel density estimation was used.

## 3. Results

The use of trait prevalence-informed priors usually had a strong impact on the performance of the prediction model, although the extent differed between EVCs and also between categories of the same EVC (see below). We found that prediction performance of prior-free models could be improved by a substantial proportion of tupels of prior values in the respective models. On the other hand, and perhaps not surprisingly, a substantial proportion of prior tupels led to a deteriorated prediction performance compared to the respective prior-free model.

### *3.1. Impact of trait prevalence-informed priors on sensitivity and specificity*

With few exceptions, sensitivity (Fig. 1) and specificity (Fig. 2) were strongly affected by variation in trait prevalence-informed prior values. A particular choice of priors could shift sensitivity usually in both directions from that of the prior-free model, often even approaching the extreme values of 0 or 1, respectively. All traits showed a strong dependence of their prediction sensitivity on the choice of prior values, most strongly for blue and intermediate eye color, blond and brown hair color, hair structure and freckles. Skin color categories seemed to be less affected by the choice of prior values especially when the darkest categories were merged, but not when the palest categories were merged. Notable exceptions were dark and dark to black skin color, which appeared barely affected by the choice of prior values. In general, specificity of predicting lighter eye and skin color was more strongly impacted by changing prior values than darker tones, as were straight and wavy hair structure categories as well as the presence of freckles. Similarly, blond and brown hair colors were more strongly affected compared to the categories of red and black hair color. Strikingly, dark skin and hair color, but also red hair and curly hair structure appeared almost insensitive in their prediction specificity when it comes to the use of prevalence priors. Interestingly, the probability of a shift away from the prior-free prediction differed between the directions as well as the average extent of this shift for both sensitivity (Table 2) and specificity (Table 3) across all EVCs. In general, we noticed that most of the prior tupels were above or equal to the prior- free value for all EVCs apart from a few exceptions. These exceptions included some skin color categories such as very pale and intermediate, whose sensitivity seemed to be lower than that of the prior-free approach for most prior tupels. The majority of prior tupels for the specificity of blue and intermediate eye color also resulted into lower values than the prior-free approach.

Of note, the distributions of sensitivity and specificity across the space of possible prior values assumed an almost discrete form for skin color when the darkest categories merged, most prominently for the light skin categories. Predicting dark and dark to black skin colors by using prevalence priors does not show any difference from the performance of the prior-free approach, also in the case where these two

**Table 2**
Shift in sensitivity in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Category | Below [%] | Above [%] |
|---|---|---|---|
| **Eye color** | Blue | 35.6 | 64.4 |
| | Intermediate | 38.6 | 61.4 |
| | Brown | 28.9 | 71.1 |
| **Hair color** | Blond | 51.4 | 48.6 |
| | Brown | 49.8 | 50.1 |
| | Red | 56.6 | 43.4 |
| | Black | 0.0 | 100.0 |
| **Skin color (4/5)** | Very Pale | 78.9 | 21.1 |
| | Pale | 55.7 | 44.3 |
| | Intermediate | 63.1 | 36.9 |
| | Dark/Dark to Black | 0.0 | 100.0 |
| **Skin color (1/2)** | Very Pale/Pale | 48.9 | 51.1 |
| | Intermediate | 48.9 | 51.1 |
| | Dark | 0.0 | 100.0 |
| | Dark to Black | 0.0 | 100.0 |
| **Hair structure** | Straight | 38.3 | 61.7 |
| | Wavy | 59.0 | 41.0 |
| | Curly | 98.9 | 1.1 |
| **Freckles** | Freckled/Non-freckled | 49.5 | 50.5 |

Proportion of prior tupels resulting in sensitivity values below and above the value for the prior-free approach, respectively.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

**Table 3**
Shift in specificity in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Category | Below [%] | Above [%] |
|---|---|---|---|
| **Eye color** | Blue | 52.5 | 47.5 |
| | Intermediate | 60.7 | 39.3 |
| | Brown | 27.7 | 72.3 |
| **Hair color** | Blond | 52.8 | 47.2 |
| | Brown | 49.4 | 50.6 |
| | Red | 43.5 | 56.5 |
| | Black | 24.9 | 75.1 |
| **Skin color (4/5)** | Very Pale | 52.8 | 47.2 |
| | Pale | 53.8 | 46.2 |
| | Intermediate | 41.2 | 58.8 |
| | Dark/Dark to Black | 100.0 | 0.0 |
| **Skin color (1/2)** | Very Pale/Pale | 48.9 | 51.1 |
| | Intermediate | 48.9 | 51.1 |
| | Dark | 0.0 | 100.0 |
| | Dark to Black | 0.0 | 100.0 |
| **Hair structure** | Straight | 59.0 | 41.0 |
| | Wavy | 39.7 | 60.3 |
| | Curly | 41.6 | 58.4 |
| **Freckles** | Freckled/Non-freckled | 49.5 | 50.5 |

Proportion of prior tupels resulting in specificity values below and above the value for the prior-free approach, respectively.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

categories were considered as a single one.

### 3.2. Impact of trait prevalence-informed priors on positive and negative predictive values

Similar to the results for sensitivity and specificity, positive predictive values (PPV; Fig. 3) and negative predictive values (NPV; Fig. 4) were, with few exceptions, strongly affected by the choice of prior values for EVCs such as eye and hair color. Quite similarly, the impact was again strongest for freckles and hair structure. More specifically for the latter, PPV appeared to be quite sensitive for all categories in the change of prior values, while the impact on NPV seems to be larger for all categories apart from curly hair. Regarding skin color, the impact of prevalence priors on PPV and NPV was very small when the darkest categories were merged. When merging the palest categories, the impact of different prior values was very small only for the categories of dark and dark to black.

The values of PPV and NPV differed regarding the direction and also

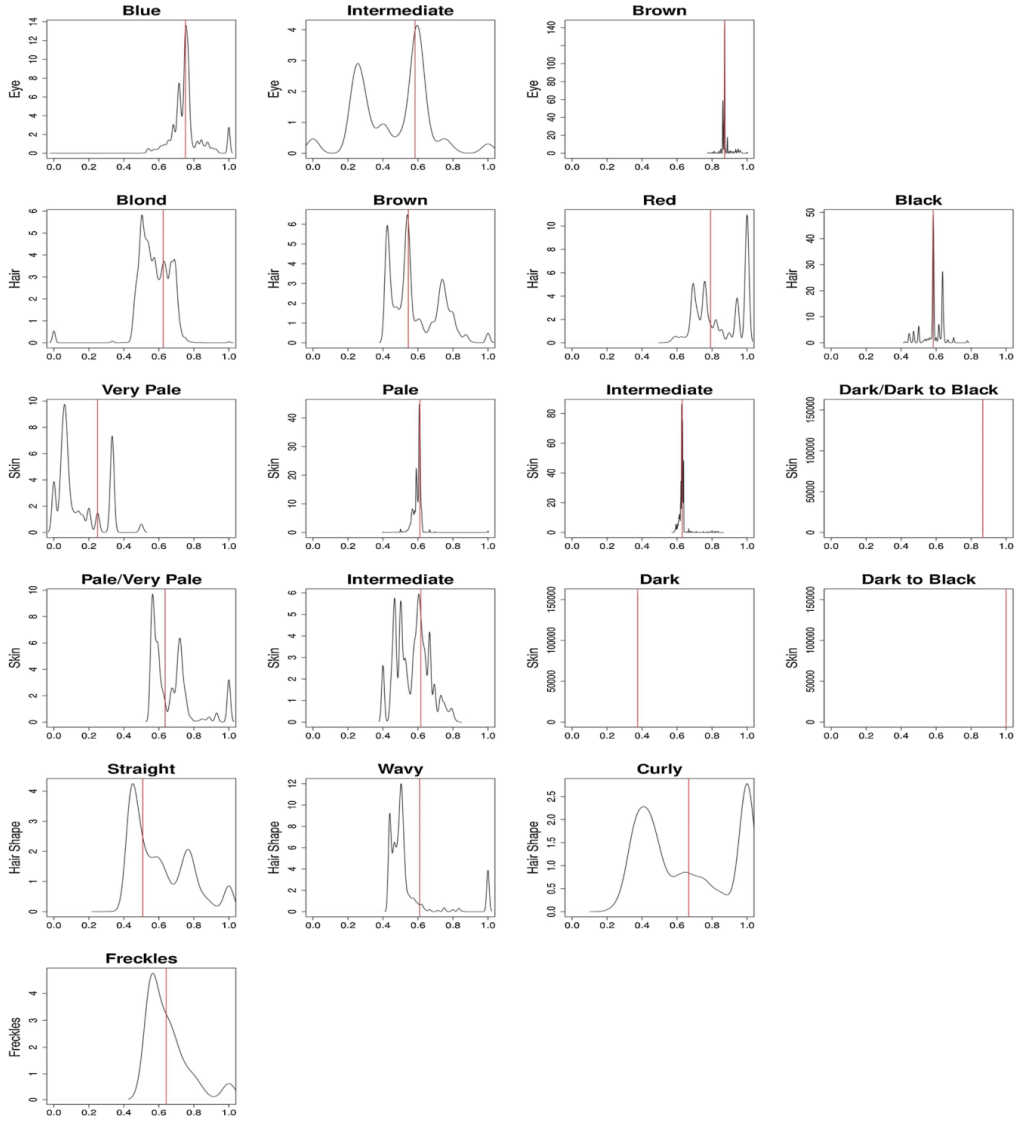


**Fig. 3. Impact of the choice of trait prevalence priors on positive predictive values (PPV) in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

the average extent for each of the considered traits when priors were incorporated in the model (Tables 4 and 5). For example, for freckles we observed that most of the prior tupels incorporated in the model seem to perform better compared to the prior free approach for both PPV and NPV. However, for most of the other traits, we observed that only almost half of the prior tupels showed a better performance when compared to the model without priors, while the other half seemed to show an inferior performance but only for specific categories with respect to both measurements. For example, the brown eye color category showed a high percentage above or equal to the prior-free value for PPV as well as NPV, while for blue and intermediate eye categories the percentage above or equal the prior-free approach is ranging around 50 %.

Red and black hair color appeared to be barely affected by the choice of prior tupels compared to blond and brown especially for NPV, while freckles and eye color showed in general high susceptibility in both measurements, with the only exception of the brown eye color category which seemed less impacted. Generally, we observed small effects for skin color when dark and dark to black categories were considered as

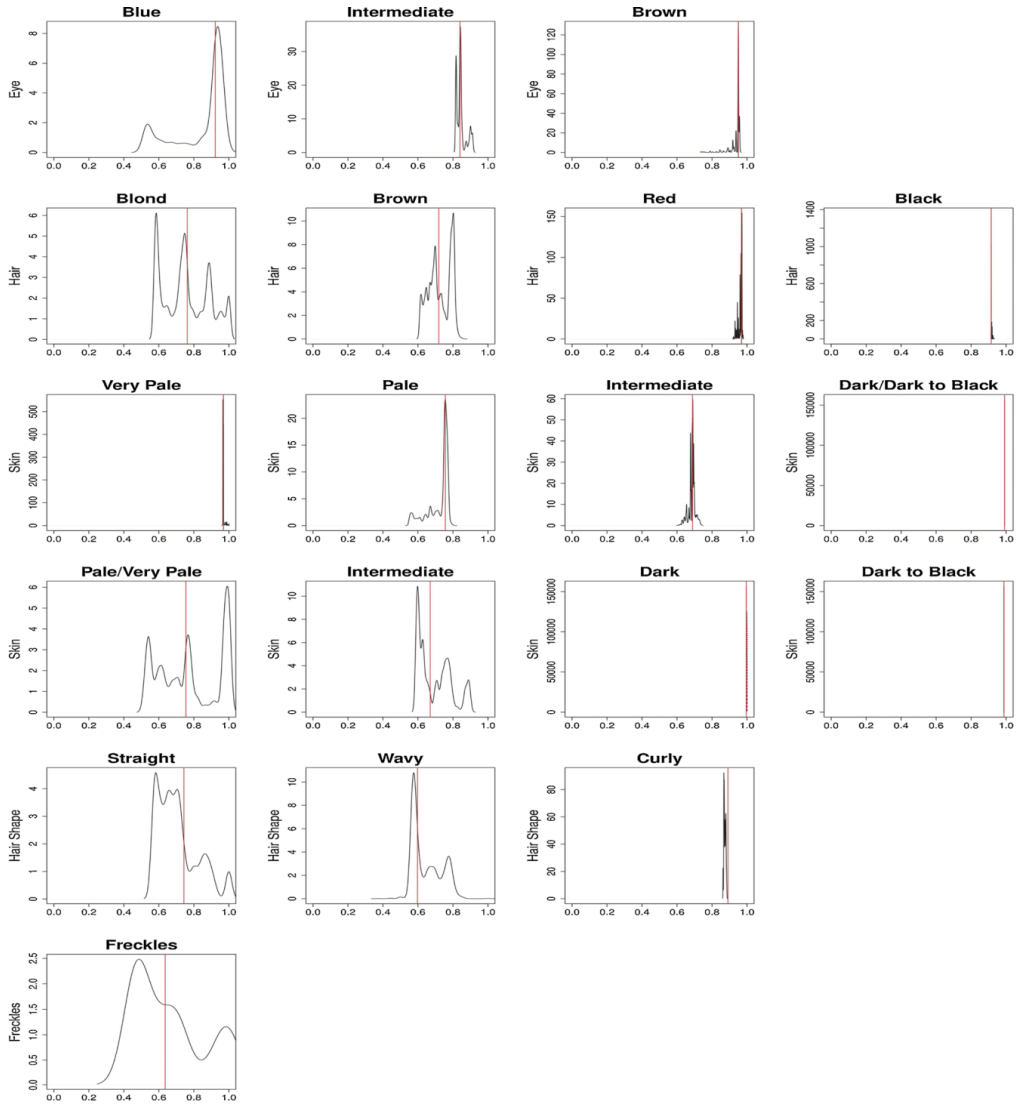


**Fig. 4. Impact of the choice of trait prevalence priors on negative predictive values (NPV) in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

**Table 4**
Shift in PPV in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Category | Below [%] | Above [%] |
|---|---|---|---|
| **Eye color** | Blue | 54.5 | 41.7 |
| | Intermediate | 59.3 | 24.3 |
| | Brown | 37.4 | 62.6 |
| **Hair color** | Blond | 63.2 | 30.3 |
| | Brown | 49.8 | 50.2 |
| | Red | 45.3 | 54.7 |
| | Black | 60.9 | 39.1 |
| **Skin color (4/5)** | Very Pale | 60.5 | 16.6 |
| | Pale | 82.1 | 17.9 |
| | Intermediate | 52.8 | 47.2 |
| | Dark/Dark to Black | 0.0 | 100.0 |
| **Skin color (1/2)** | Very Pale/Pale | 50.8 | 49.2 |
| | Intermediate | 69.7 | 30.3 |
| | Dark | 0.0 | 100.0 |
| | Dark to Black | 0.0 | 100.0 |
| **Hair structure** | Straight | 37.0 | 56.1 |
| | Wavy | 83.4 | 10.5 |
| | Curly | 50.9 | 44.1 |
| **Freckles** | Freckled/Non-freckled | 33.3 | 50.5 |

Proportion of prior tupels resulting in positive-predictive values (PPV) values below and above the value for the prior-free approach, respectively. In cases where percentages above and below the prior-free approach do not sum to 100 is obtained due to the occurrence of NAs in this model measurements. Thus, those observations were omitted.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

one combined category, while merging pale and very pale categories resulted in being more sensitive to the choice of the prior values. For hair structure, the NPV of the curly category was slightly impacted, while the PPV seemed to be very sensitive to prior tupel choice.

### *3.3. Impact of trait prevalence-informed priors on AUC and overall accuracy*

Finally, we assessed the overall performance by means of area-under-curve (AUC) and overall accuracy values. We generally observed only a small impact of the choice of prior values on AUC for all EVCs tested (Fig. 5). More specifically, all categories for hair structure and hair color appeared to be barely affected in AUC by the varying prior tupels, with brown, red and black showing a smaller impact compared to blond. Blue and brown eye colors were also barely affected, while intermediate eye color appeared a bit more susceptible. Similar to the aforementioned EVCs, the effect of priors on skin color categories was generally small, either when the palest either when the darkest categories were merged. The category of pale/very pale skin color appeared to perform worse in the model when priors were incorporated compared to the prior-free approach. AUC for freckles showed independence from the choice of prior values since its value remained stable for all possible prior tupels.

Regarding AUC values, most of the categories showed that almost half of the prior tupels performed above or equal to the prior-free approach (Table 6). There were few exceptions, such as hair structure and freckles, where most of the proportions were above the prior-free AUC value. Regarding very pale and pale/very pale, almost all prior prediction values seem to perform worse than the prior-free approach.

In comparison to AUC, overall prediction accuracy (Fig. 6) was much more affected by the choice of prior values. All five EVCs showed substantial susceptibility to the choice of priors reflected in the overall accuracy. Notably, there was some room for improvement for overall prediction accuracy except from hair structure, which seemed to perform worse compared to the prior-free approach. However, the overwhelming majority of prior tupels led to accuracy deterioration (Table 7). We also noticed that misspecification of priors often caused a deterioration in the prediction performance measurements for some traits as well as in the overall accuracy (Fig. 6).

**Table 5**
Shift in NPV in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Category | Below [%] | Above [%] |
|---|---|---|---|
| **Eye color** | Blue | 49.1 | 50.9 |
| | Intermediate | 53.1 | 46.9 |
| | Brown | 41.5 | 58.5 |
| **Hair color** | Blond | 58.5 | 41.5 |
| | Brown | 50.0 | 50.0 |
| | Red | 70.6 | 29.4 |
| | Black | 44.5 | 55.5 |
| **Skin color (4/5)** | Very Pale | 64.9 | 35.1 |
| | Pale | 64.3 | 35.7 |
| | Intermediate | 54.3 | 45.7 |
| | Dark/Dark to Black | 100.0 | 0.0 |
| **Skin color (1/2)** | Very Pale/Pale | 45.8 | 54.2 |
| | Intermediate | 50.8 | 49.2 |
| | Dark | 0.0 | 100.0 |
| | Dark to Black | 0.0 | 100.0 |
| **Hair structure** | Straight | 71.3 | 27.7 |
| | Wavy | 45.6 | 54.4 |
| | Curly | 100.0 | 0.0 |
| **Freckles** | Freckled/Non-freckled | 42.4 | 50.5 |

Proportion of prior tupels resulting in negative-predictive values (NPV) below and above the value for the prior-free approach, respectively.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

## 4. Discussion

In the present study, we aimed at assessing the impact of using trait prevalence-informed priors on the prediction accuracy of an expanded set of EVCs, including eye, hair and skin color as well as hair structure and freckles. Our study was motivated by the question if such prior information, possibly representing trait class prevalence in biogeographic ancestry groups, may improve the prediction accuracy of traits over prior-free models. For all EVCs except freckles, we used for our models the same predictive markers as applied in the previously established prediction models [9,11,12,15]. Although due to data availability issues the number of predictors was lower in our freckles prediction modeling than previously [13] this discrepancy shall not affect our main outcomes for freckles significantly, since we applied the same reduced marker set to both the model with and without priors. Regarding the prior information, we surprisingly noticed that there is a limited spatial and population-specific trait prevalence information available for hair, skin and eye color, hair structure [24] and even non-existent for other traits such as freckles. We therefore exhaustively investigated the impact of the choice of prior values for the different trait categories on a fine-grained grid of all possible sets, or tupels, of values to obtain a general picture of the impact of priors on prediction performance. To this end, we trained and tested Bayesian versions of multinomial logistic regression (MLR) and binomial logistic regression (BLR) models, respectively, and compared their performance to the respective prior-free versions, using different trait-specific data sets.

Our results showed that the use of trait prevalence-informed priors can have a strong impact on the performance of the prediction models for the 5 EVCs tested. Such use carries some potential to improve the prediction of most EVCs and some of their categories compared to a prior-free approach, as evidenced by a substantial proportion of prior tupels with better performance statistics. However, we also found large proportions of prior tupels that led to inferior prediction results, indicating the risk that the misspecification of those priors may lead to a gross deterioration in the model performance. This deterioration could be explained by the fact that the true prevalence values are unknown. The prior-free approach is influenced by the proportions of the

categories in the data set. Random splitting into separate training (80 %) and test (20 %) datasets, as performed here for all EVCs, resulted in approximately equal proportions for each of the trait categories in these two data sets, respectively. In consequence, the trained model was well adapted to the category proportions in the test data set, possibly leading to some over-fitting of the model. This may have led to a slight over-estimation of the performance of the prior-free models. Accurate trait prevalence specification is of utmost importance to obtain reliable and accurate predictions. However, with the lack of such information, the application of prior-incorporating Bayesian approaches for EVC prediction in forensic cases appears not feasible at this stage.

Given the lack of spatial or population-specific prevalence information for the EVCs considered in this study, which represented a significant obstacle to our analysis, we were not able to compare the performance of prior-incorporating and prior-free approaches against a gold standard. As gold standard we should have had reliable population-representative prior values for all EVCs and their categories, which, however, are not available. Therefore, we explored the impact of priors across the whole space of

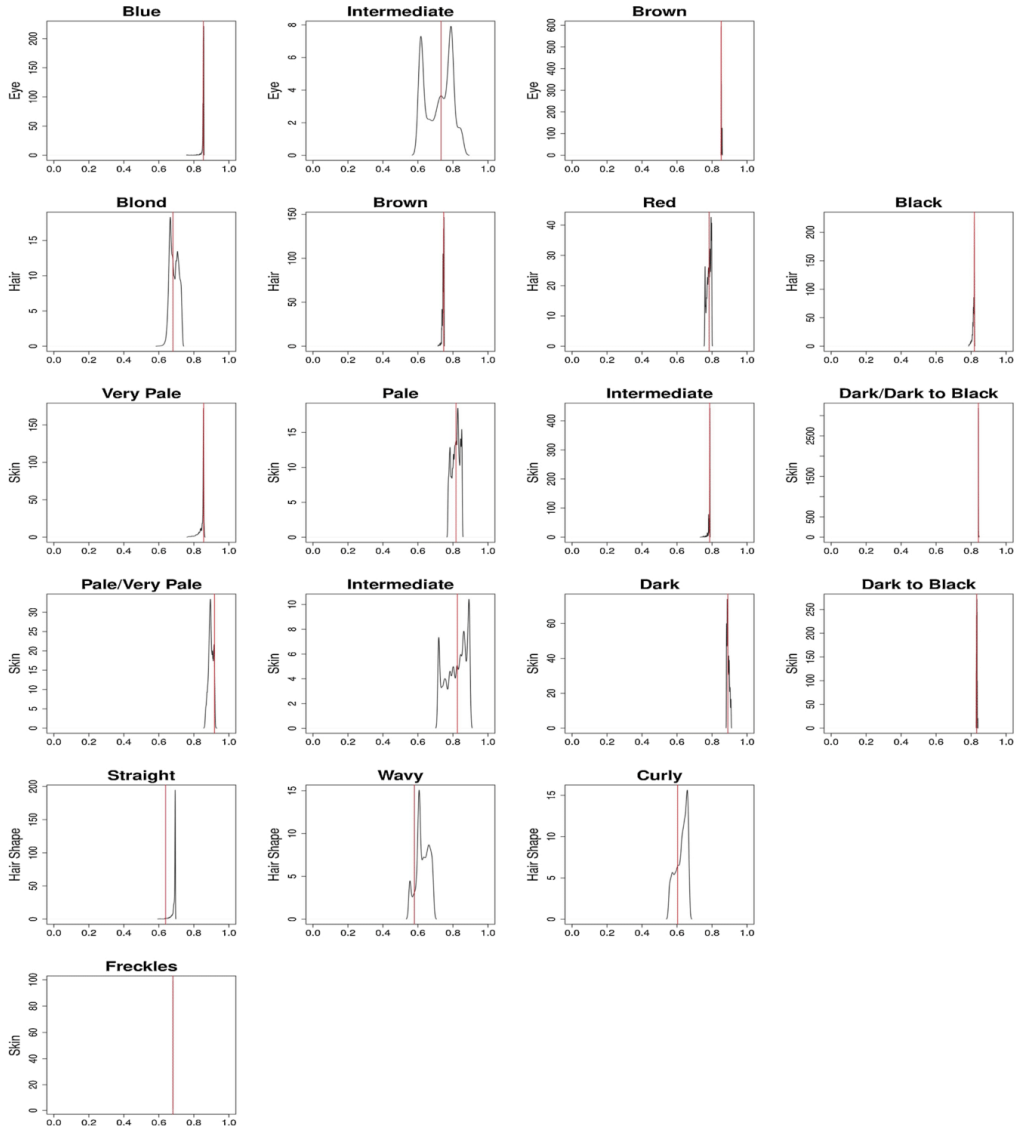


**Fig. 5. Impact of the choice of trait prevalence priors on the area-under-curve (AUC) in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

possible tupels. Another possible interpretation of our approach, given the lack of knowledge about the underlying "truth" regarding the knowledge on trait prevalence over geographic space, is that the priors resemble differential costs for misclassification, which may also be an interesting future approach in forensic applications.

Little susceptibility of the prediction outcome to the choice of prior values, represented by likelihood ratio values of large magnitude compared to those of the priors, likely reflects a large extent of genetic determination of a trait or a particular trait category and that a large proportion of the causal genetic variants determining this trait, or at least their strongly correlated proxies, are already included in the prediction model [5,37–39]. This agrees with the statement of Caliebe et al. [33] that trait prevalence values provide no (or little) additional information if all (or almost all) genetic trait-determining variants are included as predictors in the model, i.e. that the prediction is independent of the population. From all EVCs and their categories investigated here, red hair color prediction comes closest to this, as red hair is determined by only one gene, *MC1R*, from which multiple DNA variants, most of them being non-synonymous DNA variants that are likely causal, are included in the hair color prediction model based on the HIrisPlex markers for hair color prediction used here. For complex traits or trait categories, however, dozens or even hundreds of genetic factors will contribute to the trait and usually only a fraction of them is known and included in the prediction model. It is assumed that all EVCs and EVC categories, including those tested here besides red hair are complex traits or trait categories determined by large numbers of genes, respectively. This was already demonstrated for hair and skin color based on large-scale genome-wide association studies (GWAS) [40,41], and therefore is also expected for eye color for which such a large-scale GWAS is currently pending. For hair shape and freckles the previous GWAS were not yet on such large scale, but those multiple genes that were successfully identified showed mostly small effect sizes and explained only a fraction of the estimated heritability [42,43]; only large-size GWAS will be able to increase the explained heritably in the future. For complex phenotypes, use of prevalence values may actually increase prediction accuracy if specified correctly, because they contain information on, and can act as proxies for, those variants that also contribute but are not included in the model.

The strong dependency of prediction performance on priors for most traits and categories further reflects that many, if not most, predictions are made based on only moderately different posterior probabilities and, in turn, likelihoods do not differ strongly between the categories, because not all causal factors are yet known and could therefore be included in the prediction models. Use of priors may then easily shift classification decisions, thereby simply facilitating a trade-off between sensitivity and specificity as well as PPV and NPV in the absence of information on true trait prevalence values. Interestingly, the AUC appeared to be largely unaffected by changing prior tupels.

Both observations, the potential for prediction improvement by use of priors as well as the risk of inferior performance when those priors are mis-specified, motivate future studies. An important and preferable way would be to identify more causal genetic factors involved in EVC etiology, thereby obliterating the need for proxies of those causal factors. However, given their likely small and at most moderate effects, this would require very large data sets for future studies to identify such genetic variants. For instance, a recent GWAS on hair color tested more than 290,000 individuals in an European discovery dataset that led to the identification of 124 associated independent genetic loci at genome-wide significance, of which 111 were novel [40]. However, most of these DNA variants will not be causal themselves, because of the focus of commonly used SNP microarrays on markers that allow for good imputation of other, common markers ('imputation backbone'), while providing only limited numbers of SNPs centered on gene regions or selected phenotypic relevance ('contents enrichment').

Another area for future research is to collect, for as many populations from as many geographic regions that are relevant based on the phenotypic variation of the EVCs to be predicted, trait prevalence data on the same or higher level of detail (e.g. categories) as achievable by DNA-based EVC prediction. However, even when such data are available, the use of forensic ancestry DNA testing to identify the geographic region for which EVC trait prevalence data are to be allocated for use as priors in EVC prediction will only be applicable, in case the prevalence values for different populations within such DNA-identified region do not show much variation, and if the regional geographic ancestry can be inferred with high confidence from the crime scene DNA sample. While collection of prevalence data may be achievable in the future, provided such studies are carried out with suitable geographic coverage and EVC phenotypic details, and given that regional such as continental ancestry inference based on enough DNA markers already is possible [44], the trait variation within DNA-identifiable geographic regions remains as problem. For instance, within Europe, which as continental region is identifiable with forensic DNA ancestry testing [44], eye and hair color prevalence values largely vary between populations from different parts of Europe. Thus, averaging such population prevalence values, if available, will not result in suitable priors for any person originating from any European population. This could only be solved by increasing the level of detail of DNA-based ancestry testing to the sub-regional or even population level, which currently, however, is not achievable and also is not expected to be achievable in the near future. Identifying genetic geographic population substructure within continents, such as within Europe [45], requires thousands of autosomal SNPs – a number that currently cannot be achieved given available technologies that are suitable for forensic DNA analysis. The simultaneous and targeted analysis of many thousands of SNPs in low-quantity and low-quality DNA typically available from crime scene stains requires the development of new DNA technology in the future.

In summary, our results provide a first assessment of the impact of trait prevalence-informed priors on the prediction model performance for several EVCs. Incorporation of priors, possibly informed by trait class prevalence values in biogeographic ancestry groups, can improve the performance of predicting appearance traits, but a correct specification of those priors appears mandatory to protect against a deteriorated performance. Future work is needed to obtain unbiased estimates of trait prevalence for EVCs to be predicted in a large variety of populations, when mostly non-causal genetic

**Table 6**
Shift in AUC in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Category | Below [%] | Above [%] |
|---|---|---|---|
| **Eye color** | Blue | 50.5 | 49.5 |
| | Intermediate | 50.5 | 49.5 |
| | Brown | 43.8 | 56.2 |
| **Hair color** | Blond | 43.9 | 56.1 |
| | Brown | 70.9 | 29.1 |
| | Red | 49.2 | 50.8 |
| | Black | 84.9 | 15.1 |
| **Skin color (4/5)** | Very Pale | 90.1 | 9.90 |
| | Pale | 49.2 | 50.8 |
| | Intermediate | 46.2 | 53.8 |
| | Dark/Dark to Black | 53.4 | 46.6 |
| **Skin color (1/2)** | Very Pale/Pale | 96.4 | 3.64 |
| | Intermediate | 50.8 | 49.2 |
| | Dark | 50.8 | 49.2 |
| | Dark to Black | 50.8 | 49.2 |
| **Hair structure** | Straight | 1.32 | 98.7 |
| | Wavy | 11.6 | 88.4 |
| | Curly | 27.8 | 72.3 |
| **Freckles** | Freckled/Non-freckled | 0.0 | 100.0 |

Proportion of prior tupels resulting in area-under-curve (AUC) values below and above the value for the prior-free approach, respectively.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

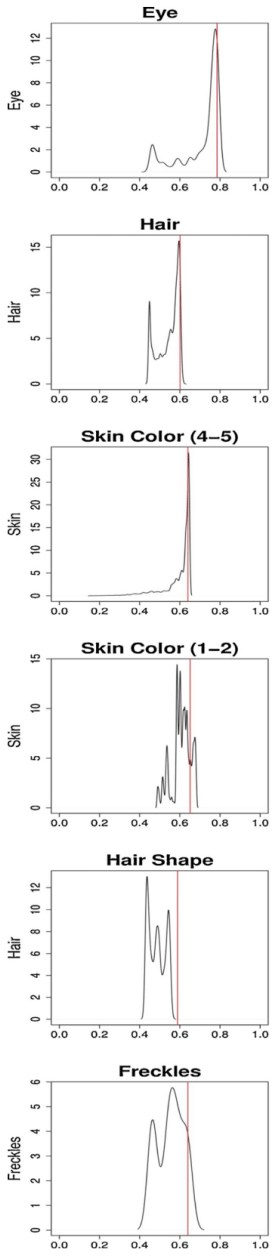


**Fig. 6. Impact of the choice of trait prevalence priors on the overall accuracy in EVC prediction modeling from genetic data.** Results are presented for a Bayesian approach using a multinomial logistic regression model for predicting four pigmentation trait categories, namely those of eye color (EC; first line), hair color (HC; second line) and skin color (SC; third line: darkest categories merged; fourth line: palest categories merged), where the vertical line corresponds to a prior-free prediction.

**Table 7**
Shift in overall accuracy in EVC prediction modeling from genetic data for the prior-based models compared to the prior-free models.

| Trait | Below [%] | Above [%] |
|---|---|---|
| **Eye color** | 75.24 | 24.75 |
| **Hair color** | 90.97 | 9.027 |
| **Skin color (4/5)** | 72.96 | 27.03 |
| **Skin color (1/2)** | 80.92 | 19.07 |
| **Hair structure** | 100.0 | 0.0 |
| **Freckles** | 87.87 | 12.12 |

Proportion of prior tupels resulting in overall accuracy values below and above the value for the prior-free approach, respectively.
Skin color (4/5) is referring to the skin color prediction when the two darkest categories of dark and dark to black were merged and considered as one single category. Similarly Skin color (1/2) is referring to the case when the two palest categories of very pale and pale were merged and considered as one.

markers are continued to being used for trait prediction. This need will be reinforced by future GWAS whose larger sample sizes will allow the detection of genetic markers with even smaller effect sizes, yet most of them likely being non-causal. Finally, appearance trait research has to overcome the assembly of ever more associated, yet non-causal genetic markers and, via experimental evidence, has to arrive at the identification of the actual causal genetic factors for EVCs. If successful, this will allow to achieve accurate EVC prediction in a population-independent way, eventually rendering the use of trait prevalence priors obsolete in the future.

## Funding

This study received support from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 740580 within the framework of the *Visible Attributes through Genomics* (VISAGE) Project and Consortium. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd and the National Institute for Health Research (NIHR) funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. The funding organization had no influence on the design, conduct or conclusions of the study.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Appendix

*Centres and investigators of the VISible Attributes through GEnomics (VISAGE) Consortium*

**Website:** http://www.visage-h2020.eu/Erasmus **University Medical Center Rotterdam (Netherlands):** Manfred Kayser, Vivian Kalamara, Arwin Ralf, Athina Vidaki.

**Jagiellonian University (Poland):** Wojciech Branicki, Ewelina Pośpiech, Aleksandra Pisarek.

**Universidade de Santiago de Compostela (Spain):** Ángel Carracedo, Maria Victoria Lareu, Christopher Phillips, Ana Freire-Aradas, Ana Mosquera-Miguel, María de la Puente.

**Medizinische Universität Innsbruck (Austria):** Walther Parson, Catarina Xavier, Antonia Heidegger, Harald Niederstätter.

**Universität zu Köln (Germany):** Michael Nothnagel, Maria-Alexandra Katsara, Tarek Khellaf.

**King's College London (United Kingdom):** Barbara Prainsack, Gabrielle Samuel.

**Klinikum der Universität zu Köln (Germany):** Peter M. Schneider,

Theresa E. Gross, Jan Fleckhaus.

**Bundeskriminalamt (Germany):** Ingo Bastisch, Nathalie Schury, Jens Teodoridis, Martina Unterländer.

**Institut National De Police Scientifique (France):** François-Xavier Laurent, Caroline Bouakaze, Yann Chantrel, Anna Delest, Clémence Hollard, Ayhan Ulus, Julien Vannier.

**Netherlands Forensic Institute (Netherlands):** Titia Sijen, Kris van der Gaag, Marina Ventayol-Garcia.

**National Forensic Centre, Swedish Police Authority (Sweden):** Johannes Hedman, Klara Junker, Maja Sidstedt.

**Metropolitan Police Service, London (United Kingdom):** Shazia Khan, Carole E. Ames, Andrew Revoir.

**Centralne Laboratorium Kryminalistyczne Policji (Poland):** Magdalena Spólnicka, Ewa Kartasińska, Anna Woźniak.

## References


[1] M. Kayser, Forensic DNA Phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.

[2] M. Kayser, d.K. P, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[3] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (3) (2009) 154–161.

[4] F. Liu, et al., Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. (2009) 19.

[5] W. Branicki et al. Model-based prediction of human hair color using DNA variants. 129(4) (2011): p. 443-454.

[6] S. Walsh, et al., Global skin colour prediction from DNA, Hum. Genet. 136 (7) (2017) 847–863.

[7] J. Alghamdi, et al., Eye color prediction using single nucleotide polymorphisms in Saudi population, Saudi J. Biol. Sci. (2018).

[8] Y. Ruiz, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2013) 28–40.

[9] S. Walsh, et al., IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information, Forensic Sci. Int. Genet. 5 (2011) 170–180.

[10] E. Pospiech, et al., The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction, Forensic Sci. Int. Genet. 11 (2014) 64–72.

[11] S. Walsh, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 7 (1) (2013) 98–115.

[12] L. Chaitanya, et al., The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation, 35, 2018, pp. 123–135.

[13] M. Kukla-Bartoszek, et al., DNA-based predictive models for the presence of freckles, Forensic Sci. Int. Genet. 42 (2019) 252–259.

[14] F. Liu, et al., Prediction of male-pattern baldness from genotypes, Eur. J. Hum. Genet. 24 (6) (2016) 895–902.

[15] E. Pospiech, et al. Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA (2018). 37: p. 241–251.

[16] B. Hernando, et al., Genetic determinants of freckle occurrence in the Spanish population: towards ephelides prediction from human DNA samples, Forensic Sci. Int. Genet. 33 (2018) 38–47.

[17] S.P. Hagenaars, et al., Genetic prediction of male pattern baldness, PLoS Genet. 13 (2) (2017).

[18] M. Marcinska, et al., Evaluation of DNA variants associated with Androgenetic Alopecia and their potential to predict male pattern baldness, PLoS Genet. 10 (5) (2015) e0127852.

[19] F. Peng et al. Genome-Wide Association Studies Identify Multiple Genetic Loci Influencing Eyebrow Color Variation in Europeans (2019). 139: p. 1601–1605.

[20] F. Liu, et al., Common DNA variants predict tall stature in Europeans, Hum 133 (5) (2013) 587–597.

[21] S. Walsh, et al., DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 6 (3) (2012) 330–340.

[22] O. Maroñas, et al., Development of a forensic skin colour predictive test, Forensic Sci. Int. Genet. 13 (2014) 34–44.

[23] J. Söchtig, et al., Exploration of SNP variants affecting hair colour predictionin Europeans, Int. J. Legal Med. 129 (5) (2015).

[24] M.A. Katsara, M. Nothnagel, True colors: a literature review on the spatial distribution of eye and hairpigmentation, Forensic Sci. Int. Genet. 39 (2019) 109–118.

[25] G. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, Inc., New Jersey, 2004.

[26] A. Moayyeri, et al., The UK adult twin registry (TwinsUK resource), Twin Res. Hum. Genet. 16 (1) (2012) 144–149.

[27] C.A. Anderson, et al., Data quality control in genetic case-control association studies, Nat. Protoc. 5 (9) (2010) 1564–1573.

[28] T.G.P. Consortium, A global reference for human genetic variation, Nat. Int. J. Sci 526 (2015) 68–74.

[29] C.C. Chang, et al., Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (1) (2015).

[30] R. Team, Integrated Development Environment for R, RStudio, 2016.

[31] R Developent Core Team: the R Project for Statistical Computing, 2018. Available from: https://www.r-project.org/.

[32] Venables, W.N. and B.D. Ripley, Modern Applied Statistics with S, 2002, New York: Springer.

[33] A. Caliebe, et al., Likelihood ratio and posterior odds in forensic genetics: two sides of the same coin, Forensic Sci. Int. Genet. 28 (2017) 203–210.

[34] Core, M.K.C.f.J.W.a.S.W.a.A.W.a.C.K.a.A.E.a.T.C.a.Z.M.a.B.K.a.t.R., caret: Classification and Regression Training (2019).

[35] J. Tuszynski, caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, 2019 etc.

[36] K. Soetaert, plot3D: Plotting Multi-Dimensional Data, 2017.

[37] J.V. Schaffer, J.L. Bolognia, The melanocortin-1 receptor: red hair and beyond, Arch. Dermatol. 137 (11) (2001) 1477–1485.

[38] C.J. Binkley, et al., Genetic variations associated with red hair color and fear of dental pain, anxiety regarding dental care and avoidance of dental care, J. Am. Dent. Assoc. 140 (7) (2009) 896–905.

[39] A. Siewierska-Gorska et al. Association of five SNPs with human hair colour in the Polish population 68(2) (2017): p. 134–144.

[40] P.G. Hysi, et al., Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability, Nat. Genet. 50 (2018) 652–656.

[41] A. Visconti, et al., Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure, Nat. Commun. (2018) 9.

[42] F. Liu, et al., Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair, Human Molecular Gentics 27 (3) (2018) 559–575.

[43] P. Sulem, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat. Genet. 40 (2008) 835–837.

[44] P.M. Schneider, et al., The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry, Dtsch. Arztebl. Int. 51-52 (51-52) (2019) 873–880.

[45] O. Lao, et al., Correlation between genetic and geographic structure in Europe, Curr. Biol. 18 (16) (2008) 1241–1248.

## 11.3 Evaluation of supervised machine-learning methods for predicting appearance traits from DNA

Submitted in Forensic Science International Genetics.

Current state: Under revision.

Maria-Alexandra Katsara[1], Wojciech Branicki[2], Susan Walsh[3], Manfred Kayser[4], Michael Nothnagel[1,5,*], on behalf of the VISAGE Consortium

[1] Cologne Center for Genomics, University of Cologne, Cologne, Germany

[2] Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

[3] Department of Biology, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN, USA

[4] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, the Netherlands

[5] University Hospital Cologne, Cologne, Germany

[*] Correspondence:

Michael Nothnagel, Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany

Tel.: +49-221-478-96847, E-mail: michael.nothnagel@uni-koeln.de

## Highlights

- Comparison of machine-learning (ML) classifiers for pigmentation trait prediction.
- Multinomial logistic regression and ML methods perform highly similar.
- ML classifiers provide no advantage with current limited marker sets.

# Abstract

The prediction of human externally visible characteristics (EVCs) based solely on DNA information has become an established approach in forensic and anthropological genetics in recent years. The main purpose of forensic DNA phenotyping is to trace individuals unknown to the investigating authorities who cannot be identified with the current comparative methods of DNA profiling. While for a large set of EVCs, predictive models have already been established using multinomial logistic regression (MLR), the prediction performances of other possible classification methods have not been thoroughly investigated thus far. Motivated by the question to identify a potential classifier that outperforms these specific trait models, we conducted a systematic comparison between MLR and three popular machine learning (ML) classifiers, namely, support vector machines (SVM), random forest (RF) and artificial neural networks (ANN). As examples, we used eye, hair and skin color categories as phenotypes and genotypes based on the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers. We compared and assessed the performances of each of the four methods, complemented by detailed hyperparameter tuning that was applied to some of the methods in order to maximize their performance. Overall, we observed that all four classification methods showed rather similar performance, with no method being substantially superior to the others for any of the traits, although performances varied slightly across the different traits and more so across the trait categories. Hence, based on our findings, none of the ML methods applied here provide any advantage on appearance prediction, at least when it comes to eye, hair, and skin color and the IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers used here.

# Introduction

In recent years, Forensic DNA Phenotyping (FDP), used to predict Externally Visible Characteristics (EVCs) of unknown crime scene sample donors or unknown deceased persons directly from DNA, has become a suitable addition to the forensic genetics toolbox. In criminal cases where suspects are unknown to the investigating authorities and therefore cannot be identified by comparative forensic DNA profiling, FDP can be used to generate investigative leads to help find unknown suspected perpetrators, and can also help in missing person identification when known relatives or ante mortem samples are not available [1-3]. By using FDP outcomes, police investigators can narrow down a large number of potential suspects, as is the case without known suspects, and they can subsequently proceed to generate standard forensic STR profiles for a reduced set of individuals that visually share such EVC FDP predicted outcomes.

As a prerequisite for developing FDP markers, in the past decade many studies have identified genetic markers involved in pigmentation traits [4-11]. Moreover, other studies have used them for developing lab tools and statistical tools for predicting eye, hair and skin color through DNA markers [12-20]. Most widely used predictive marker sets, lab tools and statistical models include in the IrisPlex system [13, 17, 21] for eye color prediction, the HIrisPlex system [20] for hair (and eye) color prediction, and the HIrisPlex-S system [19] for skin (and hair and eye) color prediction. The aforementioned statistical models are based on multinomial logistic regression (MLR) using established genetic marker panels, resulting in posterior probabilities for each trait category i.e., three eye color, four hair color, and five skin color categories [19], and are publically available for use via https://hirisplex.erasmusmc.nl/. Almost all previously established pigmentation prediction models were based on MLR. Some exceptions include fuzzy logic, artificial neural networks and classification trees used by Liu et al. [13] for eye colour prediction modelling and Snipper [14], which is a Bayesian classifier that provides the prediction results as likelihood ratios. Further exceptions include the iterative naïve Bayesian approach from Maroñas and Söchtig [22, 23] for skin and hair color respectively, and classification trees and partition modeling applied by Allwood et al. [24] (see [25] for a further review).

Currently, machine learning (ML) has become a powerful and widely used method for solving classification and clustering problems. It is a field in data analytics that focuses on the development of mathematical models that have the ability to recognize patterns in the datasets and use this information to predict future events. In parts inspired by the human brain, these algorithms can be trained on the data (training data) [26]. The training data is actually a set of examples which are used in order to fit, or estimate, the parameters of the model. The use of these algorithms is motivated by problems with large numbers of classes, linear and non-linear boundaries between them and can be implemented for different applications in versatile areas such as such as those observed in medicine, education, robotics and many others [27-29]. These boundaries refer to the decision boundaries, a hyper-surface that separates the vector space in mutually exclusive sets, one for each class. They can be either straight lines or non-linear curves. Some indicative examples of ML algorithms are linear and logistic regression [30], decision trees, random forests (RF) [31], k-nearest neighbors (k-NN) [32], support-vector machines (SVM) [33] and artificial neural networks (ANN) [34]. Despite the fact that these methods have huge potential in different fields, and an ability to handle various types of data, selecting a ML algorithm for specific data sets (problems) as well as their

optimal hyperparameters to gain maximal performance can be challenging. A comparative analysis is often necessary in order to arrive at a method that provides the best prediction accuracy for the data set used.

In the context of forensic sciences, various classifiers have been used and compared for different purposes, such as the inference of biogeographic ancestry from DNA, file type detection - the identification of evidential files that criminals hide in order to mislead police authorities, glass identification etc. [35-40]. To the best of our knowledge, a systematic quantitative comparative performance analysis of different classification methods for DNA-based prediction of different appearance traits has not been conducted thus far, except for some Naïve Bayes approaches [14, 16]. In this study, we focused on the evaluation and comparison of three different popular ML approaches, namely SVM, RF and ANN, and compared them with MLR, for the set of EVCs most widely used in FDP, namely eye, hair and skin color and by using the previously established DNA predictors from the IrisPlex, HIrisPlex, and HIrisPlex-S systems. These methods were applied and results were compared according to two different datasets, namely one containing samples from different continental ancestries and one including only the European samples thereof. For all four methods, we assessed the standard performance for each trait category, and for each trait overall, with the aim to investigate whether ML is superior, or not, over conventionally used MLR for DNA-based appearance prediction using pigmentation traits as examples.

# Materials and Methods

***Data sets***

For the present study, part of the previously used datasets for the establishment of IrisPlex model for eye color [17], the HIrisPlex model for hair color [20], and the HIrisPlex-S model for skin color [19] were applied for the prediction of those EVCs. More specifically, we used phenotype and genotype datasets from 1095 samples for eye, 1702 for hair, and 1318 for skin color prediction (*complete dataset;* **CD**), originating from Europeans, Americans, South and East Asians, African, Middle Eastern and few admixed samples. Furthermore, we used the *European subset* (**ES**) of this collection in order to restrict the analysis to a more homogenous population, comprising 821 samples for eye, 1429 for hair, and 980 for skin color prediction and originating from Ireland, Poland, Russia, Germany and Spain. Samples from which these data were previously obtained had been collected for the purpose of appearance genetic research under written informed consent, and sample collections were approved by the Ethics Committee of the Jagiellonian University (KBET/17/B/2005), the Commission on Bioethics of the Regional Board of Medical Doctors in Krakow (48 KBL/OIL/2008), the Clinical Research Ethics Committee of the Cork Teaching Hospitals (ref ECM 4 (dd) 11/01/11 ) and by the Indiana University Ethical Institutional Review Board (#1409306349). These datasets were randomly split into 80% for model training and 20% for model evaluation (Table 1) for all four methods (see below).

As previously described in detail [17, 19, 20], eye colour was classified into three categories (blue, intermediate, brown) and hair colour into four categories (red, blond, brown, black), while skin colour was classified into five categories (very pale, pale, intermediate, dark, dark to black),

following previously established categories. Since the European subset did not comprise samples with dark or dark to black skin colour, analyses in this subset were based on three categories only (very pale, pale, intermediate). The 41 HIrisPlex-S DNA markers were previously described by Chaitanya et al. [19]. In brief, for eye colour, hair colour, and skin colour, we applied the 6 SNPs from the previously established IrisPlex model for eye color prediction [17]; the 22 SNPs used for hair color prediction from the previously reported HIrisPlex model [20], and the 36 SNPs applied for the skin color prediction from the previously described HIrisPlex-S model [19], respectively.

*Appearance trait categories*

Trait categories were coded as *categorical* variables and ascendingly named as '1', '2', '3' etc. up to the corresponding number of categories for each trait:

- Eye color: Blue (1), Intermediate (2), Brown (3)
- Hair color: Blond (1), Brown (2), Red (3), Black (4)
- Skin color: Very Pale (1), Pale (2), Intermediate (3), Dark (4), Dark to Black (5); the latter two were considered only for the complete dataset

Total samples of each color category for each trait are described in detail in Supplementary Table S1. The genetic markers included in the model were converted from their initial form of the bases adenine (A), cytosine (C), guanine (G) and thymine (T) and coded *numerically* as 0, 1, 2 where 0 indicates homozygosity of the major allele, 1 heterozygosity and 2 homozygosity of the minor allele. For example, for an autosomal marker with major allele C and minor allele T, an individual's genotype CC, CT and TT would be converted to 0, 1 and 2, respectively. In all models no interaction terms were taken into account, thus only the additive effects of the corresponding genetic markers were included, similar to the previously established models [17, 19, 20]. Given the simple nature of our data and their final coding form as described above, we did not pursue feature engineering, such as considering squared variables or their products, since this would most likely not strongly affect our final outcomes. All data sets were previously quality controlled [17, 19, 20], including deviations from Hardy-Weinberg equilibrium, excessive heterozygosity, low minor allele frequencies, genetic outlier detection using principal-components analysis etc., and could therefore be directly used for prediction modelling. Samples with missing genotype data were excluded from our analysis.

*Statistical Analysis*

The analysis was conducted in R version 3.4.3 [41] and 'RStudio' version 3.5.1 [42] using the packages 'nnet' [43], 'caret' [44], 'e1071' [45] and 'randomForest' [46]. Samples with missing genotype information were excluded.

*Classification algorithms and hyperparameter tuning*

We conducted a comparative statistical analysis in order to obtain the efficacy and classification accuracy of four different classification methods, namely Multinomial Logistic Regression (MLR), Support Vector Machines (SVM), Random Forest (RF) and Artificial Neural Networks (ANN). Tuned hyperparameters play an important role in obtaining the optimal performance and accuracy results when using SVM, RF and ANN. Each classifier requires different tuning steps and hyperparameters that need tuning and tuned values depend each time on the training dataset. For each classifier, we tested a series of values for the tuning process with the optimal hyperparameters determined based on the lower out-of-bag (OOB) prediction error. OOB is an estimation that measures the prediction error of each method. The classified results based on the optimal set of hyperparameters were used afterwards for the comparison of all classifiers. In order to assess the accuracy of classification performances, we report metrics such as sensitivity, specificity, positive predictive value, negative predictive value, area under curve, confusion matrix and overall accuracy were reported.

*Multinomial Logistic Regression (MLR)*

The MLR approach is a classification method that is used to predict a nominal dependent variable based on multiple independent variables. The independent variables can be either continuous or dichotomous. It is a simple extension of the binary logistic regression that allows the dependent variable to have more than two categories. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation in order to evaluate the probability of each category. The model can be defined as follows for the 3-class traits [30]:

$$ln\left(\frac{p_2}{p_1}\right) = \alpha_2 + \sum_{j=1}^{k}\beta_2(p_2)_j x_j \quad (1)$$

$$ln\left(\frac{p_3}{p_1}\right) = \alpha_3 + \sum_{j=1}^{k}\beta_3(p_3)_j x_j \quad (2)$$

Where $\alpha_i$,$\beta_i$ $(i = 2,3)$ are the regression coefficients and $p_i$ $(i = 1,2,3)$ are denoting the probabilities for each individual sample to belong to a certain category. The latter can be calculated as follows:

$$p_2 = \frac{exp\left(a_2 + \sum_{j=1}^{k}\beta_2(p_2)_j x_j\right)}{1 + exp\left(a_2 + \sum_{j=1}^{k}\beta_2(p_2)_j x_j\right) + exp\left(a_3 + \sum_{j=1}^{k}\beta_3(p_3)_j x_j\right)} \quad (3)$$

$$p_3 = \frac{exp\left(a_3 + \sum_{j=1}^{k}\beta_3(p_3)_j x_j\right)}{1 + exp\left(a_3 + \sum_{j=1}^{k}\beta_3(p_3)_j x_j\right) + exp\left(a_2 + \sum_{j=1}^{k}\beta_2(p_2)_j x_j\right)} \quad (4)$$

$$p_1 = 1 - p_2 - p_3 \quad (5)$$

where $x_j$ is the number of minor (less frequent) allele of the $j^{th}$ SNP and $j$ is an indicator for the number of the genetic markers included for trait prediction. For this method no parameter tuning was done. Individuals were classified to the colour category with the maximum probability $p_i$ without any threshold values to be taken into account.

*Support Vector Machines (SVM)*

SVM [33] is a machine learning approach which finds the optimal hyperplane that separates the different classes with the maximum margin i.e. the maximum distance between the data points that belong to the different categories. It can solve linear or non-linear problems regarding the kernel function used each time [47]. In our case, we applied the Gaussian radial basis function (RBF) which is a widely used kernel appropriate for non-linear classification. It can be defined as follows:

$$K(X_1,X_2) = exp\ (-\gamma\|X_1 - X_2\|^2) \quad (6)$$

where $\|X_1 - X_2\|$ is the Euclidean distance between the data points $X_1, X_2$. There are two parameters that need to be tuned when using SVM classifier with RBF kernel: the parameters of cost (C) and the kernel width parameter (γ). The parameter C determines the influence of the misclassification on the objective function and γ the shape and the smoothing of the optimal hyperplane obtained. These two parameters can significantly affect the performance of an SVM model. More specifically, large C values may lead to over-fitting models while large γ could affect the shape of the hyperplane which, as a result, can affect the classification outcomes. In order to find the optimal parameters for both CD and ES, we applied the grid-searching process between ten values of γ ($2^{-5}$, $2^{-4}$, $2^{-3}$, $2^{-2}$, $2^{-1}$, $2^{0}$, $2^{1}$, $2^{2}$, $2^{3}$, $2^{4}$) and ten values of C ($2^{-2}$, $2^{-1}$, $2^{0}$, $2^{1}$, $2^{2}$, $2^{3}$, $2^{4}$, $2^{5}$, $2^{6}$, $2^{7}$). This procedure was applied for all three traits tested and the optimal values were chosen according to the lowest OOB error (Supplementary Figures S1& S4).

*Random Forest (RF)*

The RF [31] is a ML method for classification and regression tasks. It operates by constructing multiple decision trees during training and, in order to classify a new instance, each decision tree provides a classification for input data. The majority-vote classification is then chosen as the prediction. In its implementation we chose to tune two hyperparameters: the number of trees (*ntree*) and the number of features at each split (*mtry*). Several studies have already been published that focus on the appropriate number of trees for which one could obtain optimal results from the RF model. However, different opinions have been voiced during these studies. One typical example is the study of Liaw and Wiener [46] which states that larger numbers of trees provide more stable results of variable importance. On the other hand, studies such as those by Latinne et al. [48], and Hernandez-Lobato [49] found that smaller numbers of trees can also be sufficient. The study of Oshiro et al. [50] comprehensively addressed this question by applying the RF model to 29 different data sets and comparing their Area Under Curve (AUC) values. The main conclusion of this study was that the performance of an RF model does not necessarily improve when number of trees is increased, suggesting that a range between 64 and 128 trees can provide satisfactory results.

For optimal tree number (*ntree*), we checked and compared the OOB error rate for a range of 1-1000 trees and chose, separately for each trait, that number which resulted in the lowest OOB

error rate. In Supplementary Figures S2 and S5 the best values for each trait for both CD and ES are presented. For optimal *mtry* hyperparameter values, we used the default of the integer-rounded value of $\sqrt{p}$, where *p* denotes the number of variables in the model, i.e. the number of genetic markers. The corresponding mtry values for the two datasets for eye, hair and skin color therefore equaled 2, 4 and 6, respectively.

*Artificial Neural Networks (ANN)*

ANN [34, 51] is a family of approaches for classification and clustering that was inspired by the human brain in order to recognize patterns in data sets. Its history starts from the early 1940s where McCulloch and Pitts [52] wrote a paper on the functionality of human brain neurons and modeled a simple neural network by using electrical circuits. Later on 1949 Donald Hebb [53] introduced the fundamental idea of learning by supporting that neural pathways are strengthened every time that are used (Hebbian learning). In the 1950s when computers became more advanced, many ANN approaches were developed and simulated. Some examples were the approach of Farley and Clark [54], who simulated the aforementioned Hebbian Network and also the approach of Rosenblatt [55], who created the perceptron, an algorithm for pattern recognition. The interest of ANN continued also in the 1970s where Werbos [56] introduced the backpropagation algorithm that enabled the training of multi-layer networks. More recent approaches have already been established, and successfully addressed the previous challenges of deep neural networks [57-59].

The ANN consists of connected units, or nodes, called artificial neurons and these connections, just as the functionality of the human brain, can transmit signals or activate other neurons [60]. Most ANN are organized in layers and neurons, and the input data are “moving” through them only in the forward direction until some final output is obtained. Each node has its own weight which is continuously adjusted during the training procedure until data with same labels consistently yield similar output.

A number of parameters need to be tuned in order to obtain the maximum performance of the ANN model. Here, we started by tuning the number of hidden layers. At first, we looked at a range of values, starting from 1 till 10 for the hidden layers. We obtained no significant differences in the model performance for eye color prediction when we increased the number of layers. For hair and skin color prediction, we noticed some deterioration in the model performance as we increased the number of layers. Therefore, for all three traits considered here we trained our models using only one hidden layer and used the logistic function as the activation function. Other parameters that required tuning were the layer size, referring to the number of units in the hidden layer, and the decay value, acting as a regularization parameter to avoid over-fitting. Supplementary Figures S3 and S6 give the optimal values for CD and ES respectively, according to the lowest OOB error, chosen for each of the traits.

*Accuracy assessment and comparisons*

In order to compare the performance of the different classifiers we presented the model measurements evaluated on the corresponding test datasets. More specifically, for each model we calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under curve (AUC), confusion matrix and overall accuracy. Sensitivity (true positive rate) measures the proportion of the actual positive samples that are correctly identified by the model while specificity (true negative rate) refers to the proportion of the actual negative samples that were correctly identified. In addition, PPV denotes the proportion of the correct classifications among all predictions of the trait category tested each time, and NPV refers to the proportion of the correct classifications among all predictions other than the trait category of interest. AUC is a performance measure of a classification model across all possible classification thresholds while the confusion matrix describes the performance of a classification model on the test dataset for which the true values are known. Ultimately, the overall accuracy refers to the proportion of all samples that were classified correctly.

# Results

*Parameter Tuning*

For three out of the four methods applied, namely SVM, RF and ANN, we proceeded into parameter tuning for each of the two datasets and for the three traits (i.e. eye, hair and skin color) in order to obtain the optimal performance of the classifiers. The best parameters were chosen according to the lowest out-of-bag (OOB) error. For SVM, the parameters that needed to be tuned were γ and C. We found out that the optimal value for γ was 0.03125 for all three traits and for both CD and ES. The optimal C in the CD was equal to 2 for eye and skin color and equal to 16 for hair color (Supplementary Figure S1). For the ES, optimal value of C was equal to 1 for eye and skin color and equal to 8 for hair color (Supplementary Figure S4). For RF, we needed to tune the number of trees (ntree) and the optimal values for each of the traits tested. We obtained 141 trees for eye color, 713 for hair color, and 589 for skin color for CD, respectively (Supplementary Figure S2). For the ES we obtained 349 trees for eye color, 319 for hair color and 572 for skin color (Supplementary Figure S5). Regarding ANN, the parameters that needed to be tuned were the layer size and the regularization parameter of decay for avoiding over-fitting. For the size, we obtained optimal values of 2 for eye color, 6 for hair color, and 3 for skin color for the CD, while for the ES we obtained optimal values of 7 for eye and hair color and 1 for skin color (Supplementary Figures S3 & S6). For the decay in the CD, the optimal values were equal to 0.5 for hair and skin color, while for eye color it was 0.4 (Supplementary Figure S3). For the ES we obtained the optimal values for decay equal to 0.5 for eye and hair color and 0.1 for skin color (Supplementary Figure S6).

*Overall prediction accuracy*

As shown in Table 2, in terms of overall accuracy, the four classification methods performed equally well in predicting each of the three considered EVCs. For eye color and the CD, we found that MLR and ANN were able to predict the trait with an overall accuracy of 0.79, while SVM and RF performed almost at the same level with 0.78. Similarly, for the ES the highest performance was obtained with MLR and ANN (0.69), followed by SVM and RF which overall accuracy values of 0.68 and 0.67, respectively. For hair and skin color, the discrepancies among the classifiers were higher compared to eye color for both datasets. More specifically, in the CD the highest overall accuracy for hair color was obtained with MLR (0.60), while SVM and ANN performed almost equally well with accuracies of 0.57 and 0.58, respectively. The RF classifier, however, appeared to have a slightly inferior performance compared to the other classifiers, reaching the lowest overall accuracy of all classifiers at 0.55 for hair color. Similarly, for the ES the MLR had the highest performance of 0.59, followed by ANN and SVM which accuracies were equal to 0.56 and 0.55, respectively. The RF classifier appeared to have a deteriorated performance compared to the other three classifiers. Similar behavior was observed also for skin color prediction in the CD, where the MLR classifier yielded the highest performance with an accuracy of 0.63 compared to the other methods. The SVM classifier yielded an overall accuracy equal to 0.60, while RF and ANN yielded the lowest performances of 0.59 and 0.56, respectively. For the ES both MLR and SVM raised the accuracy to 0.65 for skin color, while the ANN had the lowest accuracy performance of 0.57.

*Predictive measurements*

Similar to the results of the overall accuracies, the prediction accuracy measurements for eye color presented very little to no differences between the four methods regarding blue and brown eye color, while a few deviations between the methods were seen for intermediate eye color (Table 3). For example, the sensitivity of the intermediate eye color prediction for the CD equaled 0.20 for ANN but dropped to 0.18, 0.13 and 0.15 for MLR, SVM and RF, respectively. Another example is the PPV of the intermediate eye color prediction, which obtained its highest value of 0.63 for SVM, while it dropped to 0.58 for MLR. For the ES the PPV value of intermediate eye color was raised to 0.59 for ANN while for RF it dropped to 0.42. The confusion matrices for eye color showed, for both CD and ES, small deviations among the four classifiers. Blue and brown eye colors appeared to be better predicted by the model in comparison with the intermediate eye color (Supplementary Tables S2 & S3). AUC values were at similar levels, especially for SVM, RF and ANN, while MLR slightly outperformed (Supplementary Tables S4 & S5).

For *hair color*, we also observed rather similar prediction performances for all four methods, although more pronounced differences were seen for some trait categories (Table 4) compared to eye color (Table 3). In particular, the sensitivity of Red hair color prediction in the CD reached its highest value with MLR (0.66), followed by ANN (0.58), while its value was almost halved to 0.28 for RF, and for SVM it reached 0.21 (Table 4). The sensitivity of Black hair color prediction completely dropped to zero for SVM, while its highest value was equal to 0.31 for ANN. Another example was the PPV for Black hair color, where we obtained the highest values with MLR and RF (0.58 and 0.47, respectively), while it dropped to 0.34 for ANN. We observed a similar behavior to the CD in the ES for the sensitivity of red hair color prediction where its highest values were yielded by MLR and ANN (0.69 and 0.62, respectively), while for RF and SVM the value was halved to 0.31 and 0.23, respectively. Sensitivity of black hair color dropped to zero for SVM and RF, while its highest value was obtained with MLR (0.26). PPV for black hair color reached its highest value with MLR, while it dropped to zero for RF. The confusion matrices for hair color showed similar patterns for CD and ES where the categories with fewer samples in the datasets, such as red and black hair color categories, showed higher deviations compared to blond and brown hair color (Supplementary Tables S6 & S7). AUC values for MLR outperformed for most category comparisons compared to the other ML classifiers (Supplementary Tables S2 & S3).

For *skin color*, as with hair color, we also observed uneven differences between classifiers for some predictive measurements and trait categories (Table 5). For example, in the complete dataset the sensitivity of the Very Pale skin color category prediction was 0.11 for both MLR and SVM but zero when RF and ANN were applied. Similar diminution was also observed for the sensitivity and the PPV of RF in predicting Dark skin color. RF was the only classification method where these values equaled zero (Table 5). Higher discrepancies were also observed for the specificity of pale skin color where its highest values were obtained for both MLR and RF (0.60); with SVM was applied the value dropped to 0.40. Sensitivity of dark to black category dropped to 0.66 for ANN, while for SVM and RF it reached the highest value of 0.96. In the ES, the sensitivity of very pale

skin color reached the highest value of 0.25 with MLR, while for the rest of the classifiers it was almost equal to zero. The specificity of pale skin color yielded its highest value of 0.65 with MLR but dropped to 0.40 for RF. For most of the other skin color categories and predictive measurements, the four classification methods performed almost equally (Table 5). In the confusion matrices for skin color, the categories with the highest number of samples, namely Pale and Intermediate categories, were better predicted in comparison to the other categories (Supplementary Tables S8 & S9). Also and similar to eye and hair color prediction, the AUC values for MLR mostly outperformed the other classifiers (Supplementary Tables S2 & S3).

# Discussion

In the present study, we compared four different ML classification methods, namely MLR, as widely used for EVC prediction from DNA in general, and pigmentation prediction in particular, in addition to SVM, RF and ANN with respect to their ability to predict various eye, hair and skin color categories based on the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers. The basic motivation for this study was to investigate and to identify, for each of the tested EVCs, the optimal classifier yielding the highest performance. In order to obtain the maximum performance of the SVM, RF and ANN methods, we first needed to perform hyperparameter tuning. Parameters such as cost and gamma for SVM, ntree for RF and size and decay for ANN were tuned and their optimal values were chosen according to the lowest OOB error (Supplementary Figures S1-S6).

Our results showed that when it comes to overall accuracy, all four classifiers performed almost equally well for all pigmentation traits tested, with almost no variation across the classifiers for eye color and slight variation for hair and skin color. Thus, none of the other ML methods outperformed the conventional method of MLR in predicting eye, hair and skin color based on the IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers, respectively. When looking at the full suite of prediction measurements per each of the three pigmentation traits, we noted slight differences between some classifiers for several trait categories, somewhat more for hair and skin color than for eye color. However, these differences do not allow a conclusion that any of the three ML classifiers perform superior over MLR, which is supported by our conclusion derived from the overall accuracy results. This pattern was also observed when we compared the prediction performances between the two datasets, CD and ES, where highest deviations were observed for hair and skin color compared to eye color. This was to be expected since European samples represent the major part of the CD, implying that our model was trained mostly on European samples and therefore, when we compare the performance of the CD-derived model with the one trained on the ES, we do not expect to see high differences in the overall performance.

For eye color and for both datasets, we saw a small but noticeable deviation between the four classification methods for the intermediate eye color category, while for blue and brown eye color categories, all four methods performed almost identically. As obtained with all four methods, prediction accuracies were high for blue and brown eye color, but low for intermediate eye color. This finding is in line with previous results obtained mostly based on MLR [13, 17, 19-21, 24]. As emphasized in all previous IrisPlex publications [17, 19, 20], the six IrisPlex DNA markers used here are very suitable for predicting blue and brown eye color, while their ability to predict non-blue and non-brown eye colors, which are all grouped into the intermediate eye color category, is limited. Currently, it is proposed that the limitation to predict intermediate eye color with all four classification methods is more likely explained by missing DNA predictors as opposed to the modeling type. Similarly, it may be caused by phenotype definition, as the intermediate eye color category can be expected to be more heterogeneous than the blue and brown eye colour categories that both reflect the two extremes of the eye colour phenotype distribution. A large-scale genome-wide association study (GWAS) on eye color is currently underway, aiming to increase the number of independently eye color associated DNA variants. Thus, their future use in prediction modelling of categorical eye color will help ascertain if it is the number of DNA predictors that underlies the

currently limited prediction accuracy of intermediate eye color, which based on our current findings appears to be independent of the classification method used.

Regarding hair color, the prediction performances among the four classifiers were also quite similar for the two datasets; however, the deviations were higher compared to eye color, while skin color was the trait with the highest deviations among the model measurements for some categories. This could possibly be explained by the fact that these traits and especially hair and skin color are adaptive traits that can be affected by some external or environmental factors that are not included in the genetic prediction models and consequently can affect the prediction outcomes of the different methods at various extents. In other words, each classification method has probably a different level of sensitivity in detecting such external factors, which possibly leads to higher deviations between the results. Another explanation could be the much larger number of predictors included in the hair and skin color model compared to the few markers in the eye color model, giving the ML models more freedom to pick up local patterns in the parameter space, although such patterns may represent random events that deteriorate the performance of such approaches.

The non-substantial differences obtained in the overall accuracies of the four classifiers could be explained by the fact that we only look at the additive effects of the genetic markers and not at potential interaction effects. This may be due to the underlying genetic mechanisms, but may be equally well explained by the way those genetic markers included in the established MLR models were identified in the first place. The latter has been usually done in GWAS, which mostly focuses on additive independent marker contributions to the traits. Possible incorporation of interactive effects could add some additional information that might affect the prediction performances of each classifier and probably distinguish some prediction methods that are more sensitive to the addition of interactive effects. Previous studies have already identified and incorporated SNP-SNP interactions in MLR-based modelling for eye color prediction [18, 61]. However, the previously noted predictive effects of SNP-SNP interactions were small, maybe because of the use of MLR, which requires active intervention by the analyst to consider two-way or higher-order interaction effects, whereas other ML methods often do this automatically. In our case, since with the currently available DNA predictors the interaction effects were small and no substantial differences were obtained among the four classifiers, we would not recommend interaction effects at this stage. Future ML-based pigmentation prediction studies using elongated lists of DNA predictors that already are available from large-scale GWASs for hair [62], skin color [63] and will soon be for eye color shall consider these interaction effects which might improve the overall prediction performance.

Another possible explanation for the non-substantial differences between the four classification methods could be the data sizes used for each trait and the number of samples for each trait category. Since ML methods are computational methods that ‘learn’ directly from the data, the amount of the datasets used for model training can affect the model performance. When increasing the datasets, more information regarding the patterns of each group is incorporated into the model and therefore allows the observations to be separated into the different classes more accurately. This is due to them being based on data patterns and not on weak correlations that can occur in small datasets. Thus, we could expect that this may have affected, to some extent, the prediction performances of the methods applied due to the use of these currently available datasets

that may not represent all combination patterns of alleles. This can be confirmed to some extent by our case where we noticed that prediction performance was higher when using the complete dataset in comparison with the European subset which appears to have a slightly deteriorated performance, especially for eye and hair color prediction. Larger datasets in general are often necessary and interesting to be considered for future pigmentation prediction studies, in order to release the full potential of these differing ML approaches.

In summary, our results did not show substantial differences between the four ML-based methods tested to predict appearance prediction, in particular eye, hair, and skin color using the previously established IrisPlex, HIrisPlex, and HIrisPlex-S DNA markers, respectively. Given this outcome and because of the easier interpretation of MLR with respect to the modelled function compared to other ML approaches, we suggest, at least for now, the use of the MLR as the most appropriate method for predicting appearance traits from DNA, especially with regards the three pigmentation traits used here. MLR describes a simple relationship between the inputs and the outputs, which makes the outcomes of the predictions more interpretable compared to ML methods. Contributions and feature interactions can also be easily represented by the coefficients in the MLR but require active pursuit of such interactions by the analyst, while the inner workings of SVM, RF and ANN are harder to understand and interpret, although they do offer more automated consideration of interaction terms. The latter three ML methods also do not provide a direct estimate of the importance of each feature for the model's prediction performance, although secondary, resampling-based approaches exist that may provide such an assessment. Thus, for ML methods it is harder to understand the interaction between the different features in the model.

Notably, our findings and conclusions obtained are based on a relatively small number of established DNA predictors and we did not consider interactions between them. Once more appearance DNA predictors and interactions between them have been established, it would be interesting to use them in a classifier method comparison as performed here, to find out, if the results we obtained here may have been affected by the type and number of DNA markers used, or the classification of the phenotype being predicted. However, for the time being, and with the established pigmentation DNA predictors currently available, MLR remains the preferred classification method of choice for predicting categorical pigmentation traits from DNA.

**Funding**
This study received support from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 740580 within the framework of the *Visible Attributes through Genomics* (VISAGE) Project and Consortium. The IUPUI US site was supported in part by the US National Institute of Justice (NIJ) under grant number 2014-DN-BX-K031 and 2018-DU-BX-0219. None of the funding organizations had any influence on the design, conduct, or conclusions of the study.

**Conflict of Interest**
The authors declare that they have no competing interests.

# References


1. Kayser, M. and P. de Knijf, *Improving Human Forensics Through Advances in Genetics, Genomics and Molecular Biology* Nature Reviews Genetics, 2011. **12**(3): p. 179-192.
2. Kayser, M., *Forensic DNA Phenotyping: Predicting Human Appearance From Crime Scene Material for Investigative Purposes* Forensic Science International Genetics 2015. **18**: p. 33-48.
3. Kayser, M. and P.M. Schneider, *DNA-based Prediction of Human Externally Visible Characteristics in Forensics: Motivations, Scientific Challenges, and Ethical Considerations.* Forensic Science International Genetics, 2009. **3**(3): p. 154-161.
4. Liu, F., et al., *Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up.* Human Genetics, 2015. **134**(8): p. 823-835.
5. Candille, S.I., et al., *Genome-wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations.* PLoS One, 2012. **7**(10).
6. Gerstenblith, M.R., J. Shi, and M.T. Landi, *Genome-Wide Association Studies of Pigmentation and Skin Cancer: A Review and Meta-Analysis.* Pigment Cell & Melanoma Research 2010. **23**(5): p. 587-606.
7. Sulem, P., et al., *Two newly identified genetic determinants of pigmentation in Europeans.* Nature Genetics, 2008. **40**: p. 835-837.
8. Sulem, P., et al., *Genetic determinants of hair, eye and skin pigmentation in Europeans.* Nature Genetics, 2007. **39**: p. 1443-1452.
9. Han, J., et al., *A Genome-Wide Association Study Identifies Novel Alleles Associated With Hair Color and Skin* PloS Genetics, 2008. **4**(5).
10. Rawofi, L., et al., *Genome-wide Association Study of Pigmentary Traits (Skin and Iris Color) in Individuals of East Asian Ancestry.* PeerJ, 2017. **2**(5).
11. Stokowski, R.P., et al., *A Genomewide Association Study of Skin Pigmentation in a South Asian Population.* American Journal of Human Genetics 2007. **81**(6): p. 1119-1132.
12. Alghamadi, J., et al., *Eye color prediction using single nucleotide polymorphisms in Saudi population.* Saudi Journal of Biological Sciences, 2019. **26**(7): p. 1607-1612.
13. Liu, F., et al., *Eye color and the prediction of complex phenotypes from genotypes.* Current Biology, 2009. **19**(5).
14. Ruiz, Y., et al., *Further Development of Forensic Eye Color Predictive Tests* Forensic Science International Genetics, 2013. **7**(1): p. 28-40.
15. Branicki, W., et al., *Model-based Prediction of Human Hair Color Using DNA Variants* Human Genetics, 2011. **129**(4): p. 443-454.
16. Walsh, S., et al., *Global Skin Colour Prediction From DNA* Human Genetics, 2017. **136**(7): p. 847-863.

17. Walsh, S., et al., *IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information.* Forensic Science International Genetics, 2010. **5**(3): p. 170-180.
18. Pospiech, E., et al., *The Common Occurrence of Epistasis in the Determination of Human Pigmentation and Its Impact on DNA-based Pigmentation Phenotype Prediction* Forensic Science International Genetics, 2014. **11**: p. 64-72.
19. Chaitanya, L., et al., *The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation.* Forensic Science International Genetics, 2018. **35**: p. 123-135.
20. Walsh, S., et al., *The HIrisPlex System for Simultaneous Prediction of Hair and Eye Colour From DNA* Forensic Science International Genetics, 2013. **7**(1): p. 98-115.
21. Walsh, S., et al., *DNA-based Eye Colour Prediction Across Europe With the IrisPlex System* Forensic Science International Genetics, 2012. **6**(3): p. 330-340.
22. Maronas, O., et al., *Development of a Forensic Skin Colour Predictive Test* Forensic Science International Genetics, 2014.
23. Söchtig, J., et al., *Exploration of SNP Variants Affecting Hair Colour Prediction in Europeans* International Journal of Legal Medicine, 2015. **129**(5): p. 963-975.
24. S., A.J. and S. Harbison, *SNP Model Development for the Prediction of Eye Colour in New Zealand* Forensic Science International Genetics, 2013. **7**(4): p. 444-452.
25. Katsara, M.A. and M. Nothnagel, *True Colors: A literature review on the spatial distribution of eye and hair pigmentation.* Forensic Science International Genetics, 2019. **39**: p. 109-118.
26. Alpaydin, E., *Introduction to Machine Learning*. 2004: MIT Press.
27. Kotsiantis, S.B., *Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades.* Artificial Intelligence Review 2012. **37**: p. 331-344.
28. Sidey-Gibbons, J.A.M. and C.J. Sidey-Gibbons, *Machine learning in medicine: a practical introduction.* BMC Medical Research Methodology 2019.
29. Kreuziger, J., *Application Of Machine Learning To Robotics - An Analysis* 1992.
30. Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression, Second Edition*. 2000, Canada: John Wileys & sons, Inc.
31. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**: p. 5-32.
32. Mucherino, A., P.J. Papajorgji, and P.M. Pardalos, *k-Nearest Neighbor Classification. In: Data Mining in Agriculture* 2009: Springer, New York, NY.
33. Vapnik, V.N., *The Nature of Statistical Learning Theory*. 1995, Berlin, Heidelberg: Springer-Verlag.
34. Ripley, B.D., *Neural Networks and Related Methods for Classification.* Journal of the Royal Statistical Society. Series B (Methodological), 1994. **56**(3): p. 409-456.
35. Goswami, S. and E.J. Wegman, *Comparison of Different Classification Methods on Glass Identification for Forensic Research* Journal of Statistical Science and Application 2016. **4**: p. 65-84.

36. Zadora, G., *Glass analysis for forensic purposes—a comparison of classification methods.* Journal of Chemometrics, 2007.
37. Cheung, E.Y.Y., M.E. Gahan, and D. McNevin, *Prediction of Biogeographical Ancestry From Genotype: A Comparison of Classifiers.* International Journal of Legal Medicine, 2017. **131**(4): p. 901-912.
38. Karampidis, K., E. Kavallieratou, and G. Papadourakis, *Comparison of Classification Algorithms for File Type Detection A Digital Forensics Perspective.* 2017. **56**: p. 15-20.
39. Hurtado, O.M., et al., *Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics.* PLoS One, 2016. **11**(11).
40. Toma, T.T., J.M. Dawson, and D.A. Adjeroh, *Human ancestry indentification under resource constraints -- what can one chromosome tell us about human biogeographical ancestry?* BMC Medical Genomics, 2018. **11**(5).
41. Team, R.C., *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2017.
42. Team, R. *RStudio: Integrated Development Environment for R*. 2016; Available from: http://www.rstudio.com/.
43. Venables, W.N. and B.D. Ripley, *Modern Applied Statistics with S*. Fourth ed. 2002, New York Springer.
44. Kuhn, M., *caret: Classification and Regression Training.* 2020.
45. Meyer, D., et al., *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* 2019.
46. Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* R News, 2002. **2**(3): p. 18-22.
47. Kecman, V., *Support Vector Machines – An Introduction*. Vol. 177. 2005: Springer,Berlin, Heidelberg.
48. Latinne, P., D. Olivier, and C. Decaestecker, *Limiting the Number of Trees in Random Forests.* Lect Notes Comput Sci, 2001. **2096**: p. 178-187.
49. Hernandez-Lobato, D., G. Martinez-Munoz, and A. Suarez, *How large should ensembles of classifiers be?* Pattern Recognition, 2013. **46**(5): p. 1323-1336.
50. Oshiro, T.M., P.S. Perez, and J.A. Baranauskas, *How many trees in a Random Forest?* Lecture Notes in Computer Science 2012.
51. Daniel, G.G., *Artificial Neural Network. In: Runehov A.L.C.* 2013: Springer, Dordrecht.
52. McCulloch, W. and W. Pitts, *A Logical Calculus of Ideas Immanent in Nervous Activity.* Bulletin of Mathematical Biophysics., 1943. **5**(4): p. 115-133.
53. Hebb, D.O., *The Organization of Behavior*. 1949, New York: Wiley.
54. Farley, B. and W. Clark, *Simulation of self-organizing systems by digital computer.* Transactions of the IRE Professional Group on Information Theory, 1954. **4**(4): p. 76-84.

55. Rosenblatt, F., *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, 1958. **65**(6): p. 386-408.
56. Werbos, P.J., *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* 1975.
57. Schmidhuber, J., *Learning complex, extended sequences using the principle of history compression.* Neural Computation, 1992. **4**: p. 234-242.
58. Scherer, D., A.C. Müller, and S. Behnke, *Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition.* In 20th International Conference Artificial Neural Networks (ICANN), 2010: p. 92-101.
59. Ng, A.Y., et al., *Building high-level features using large scale unsupervised learning.* 2012.
60. Kriesel, D., *A Brief Introduction to Neural Networks, available at http://www.dkriesel.com.* 2007.
61. Pospiech, E., et al., *Gene–gene interactions contribute to eye colour variation in humans.* Journal of Human Genetics, 2011. **56**: p. 447-455.
62. Hysi, P.G., et al., *Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability.* Nature Genetics, 2018. **50**: p. 652-656.
63. Visconti, A., et al., *Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure.* Nature Communications, 2018. **9**.

# 12 Discussion

The discussion is focused on the two published manuscripts and the manuscript that is currently being peer reviewed.

During the last decades, FDP which is the prediction of EVC through DNA markers has become a focus of forensic genetic research. In cases where the conventional method of DNA matching methods, such as STR profiling, cannot provide matches with already available reference profiles, FDP can result in actionable leads by narrowing down a possible large set of suspects. Certain traits, namely eye, hair and skin color may be predicted with high accuracy by only a relatively small number of genetic markers. Forensically validated tools that focus on prediction of these traits have already been developed and established. Concerning the prediction of other appearance traits, the current lack of knowledge on the genetic markers or other factors that have an impact on their phenotypic variation, is still an obstacle for their further application in the area of forensics. So far, the majority of the currently developed EVC prediction models were based on the conventional MLR approach, while only few others applied alternative ML methods. Furthermore the incorporation of the biogeographic distribution of traits as priors within the prediction models was barely demonstrated. Motivated by the fact that the existing approaches for EVC prediction still have potential for improvement and due to the aforementioned shortcomings, I aimed at assessing the impact of trait prevalence informed priors and also applied and compared various ML approaches for EVC prediction. The main aim was to address whether priors of other classification methods can provide additional information for predicting appearance traits and subsequently improve the performance of the already existing models.

For the appearance traits tested here all categories were based on established scales as defined in previous studies that focused on EVC prediction. These scales indicate well-defined categories that serve for a clear distinction among them and are appropriate for forensic purposes. More specifically, within the context of forensics, such distinction among the categories might be more beneficial for the police investigations rather than some sort of continuum scales that contain highly detailed color categories. In the latter case, the trait categories are likely closer genetically to each other and probably more difficult to be distinguished and predicted with high accuracy by the prediction models. Furthermore, they might cause difficulties in interpreting by the police authorities. For this reason, color categories i.e. for hair color are defined as blond, brown, red and black, while categories such as light/dark brown or brown with auburn shades are incorporated into the simplified scale. In a similar manner this is done for the rest of the appearance traits.

## 12.1 Limitations on assembling spatial distribution data of appearance traits among different populations

Within my PhD project I initially aimed at compiling trait prevalences among different population groups for a set of appearance traits. The main motivation was to collect these data and subsequently use it as prior information within the EVC prediction model and to investigate their final impact on the prediction accuracy. In other words, the goal was to find out whether

biogeographic-ancestry-informed priors improve the prediction accuracy for some traits, especially in cases where not all causal genetic factors have been identified and included in the current models. In order to collect information on trait prevalences, I went through old anthropological studies and also more recent ones, including both genome-wide association studies as well as candidate gene studies. After the literature review, it became obvious that there was a significant lack of available and reliable data on trait prevalences. The majority of the studies available failed to meet the criteria of providing trustworthy sources of the data and population representativeness of their samples, and were excluded from the analysis. Furthermore, I noticed that despite the fact that human pigmentation has been a topic of major interest for many decades, most of the studies were focused on the pigmentation of European populations, neglecting other population groups.

In addition, most of the studies provided prevalence data on eye and hair pigmentation, while data for other traits such as skin color were barely available. Regarding other traits such as hair structure and height there were not available data at all.

Due to lack of data availability, the scope of my project became somewhat limited, since I had to focus only on eye and hair color in mostly European and not worldwide populations. However, the limitation of data availability presented an opportunity to do a rather in-depth theoretical research on the subject.

## 12.2 Current knowledge on spatial prevalence of eye and hair color

Due to the above mentioned restrictions, the initial part of my work was the compilation of data for eye and hair color among different populations. The prevalences that could be collected for eye color comprised 16 European and Central Asian countries, while for hair color, data from only seven European countries were available.

At first, I focused on the investigation of association patterns between the two traits. For this purpose, I used the available dataset on the German population from the old anthropological study of Virchow, which comprised the largest sample of approximately 6 million individuals [168]. In this dataset, I found that there is a moderate association between eye and hair color. The highest association was between blond hair and blue eyes. These outcomes were also confirmed in the spatial maps presented in the study where high prevalences of blond hair and blue eyes were found in the northern part of Europe. Interestingly, red hair color did not show any specific pattern of association with any of the other traits.

Throughout this study, I aimed at assessing the prevalence in countries for which data were not available. For this purpose, interpolation was applied, however due to the fact that the data were quite limited, only the outcomes of the neighboring countries were estimated and presented in spatial maps. These values should be treated with caution, since they might deviate from the actual prevalence values. However, the data obtained from interpolation has its merits and represents the most recent state of knowledge on the biogeographic distribution of the traits under investigation. Further studies on the spatial distribution are necessary in order to expand the current knowledge for worldwide population groups and for an extended set of appearance traits, including skin color,

hair structure, freckles, height and male pattern baldness among others. This is of particular importance for completing the picture of spatial distribution of traits, which can be used as priors for testing their actual impact on EVC prediction.

## 12.3 Impact of priors on EVC prediction

According to the aforementioned restrictions, namely the lack of reliable data, the attempt to apply and test the actual impact of priors of EVC prediction was deemed to be infeasible at this stage. Motivated by the question whether such prior information has potential to improve the prediction accuracy of the previously established models, I exhaustively investigated the impact of priors defined on a fine-grained grid that contained all its possible values from 0.01 to 0.99. Therefore, I investigated the general behavior of priors on all possible values and tested the impact in case of misspecification. During this study I used a set of specific appearance traits that included eye, hair, skin color, hair structure, and freckles. Prior incorporated models were compared with the prior-free models. The models applied here were based on multinomial logistic regression (MLR) or binomial logistic regression (BLR), depending on the number of the trait tested each time.

According to my results, trait prevalence has an impact on the performance of the EVC prediction, at least for the traits tested here and for the selected set of markers. More specifically, I found out that for all traits, there is a certain proportion of priors that seems to outperform the prior-free approaches. The impact differed regarding each trait category, since certain categories appeared to be more susceptible in prior changes compared to others. Importantly, it also turned out that there was a significant proportion of values that yielded worse performance than the prior-free models. This indicates that there is a risk of misspecifying the priors, which can lead to deterioration of the model performance. Therefore it is critical to apply accurate trait prevalence values in order to see the actual impact on EVC prediction.

Given the lack of knowledge on these priors, I investigated not only the possible benefit that priors may have in predicting EVCs but also the cost of their misspecification. This provided a first important insight on the incorporation of priors, which hopefully will motivate further studies that through the collection of accurate prevalence values can discover the possible benefits of this approach on appearance trait prediction.

Within the different traits tested here it was obtained that some trait categories showed lower susceptibility to prior variations. This likely reflects the large extent of genetic determination of these traits that is already included in the model through the predictive markers. Such example is the red hair color category which is determined mainly by one gene, the *MC1R*. Its DNA variants are acting as causal factors and affect the type of melanin produced in the human body and lead to this certain phenotype when pheomelanin is produced. For all other traits and their categories, especially hair and skin color, several GWAS studies have demonstrated their phenotypes are complex and is affected by large numbers of genes [169, 170], while GWAS for other traits such as hair structure and freckles were not in such a large scale so far. For the last two traits several genes have been identified but they appear to have only small effects on their prediction. This indicates

the need for larger studies that might identify more markers that contribute on their phenotypic variations.

The use of priors, if specified correctly, and especially for genetically complex traits might be beneficial and improve the prediction accuracy of the already existing approaches. This might be because they can incorporate information into the model that either reflects the effect of external or environmental factors that so far cannot be incorporated in the model, or the effect of additional variants that are not yet identified but show variation in their spatial distribution. In any case, both observations, namely the potential to improve prediction accuracy and the cost of misspecification, are two important findings that motivate further studies. One important approach for further studies, would be to conduct larger scale GWAS in order to identify more causal genetic factors that contribute to the different phenotypes. Due to the fact that these factors might have small or even moderate effects, large studies are necessary to identify genetic variants. Another issue that is highlighted throughout this study, is the collection of population-based trait prevalence values on different EVCs that are to be predicted. Such data will be useful in order to obtain the actual impact on EVC prediction, which currently shows potential for improvement. In forensic cases, such priors will only be applicable when biogeographic ancestry of the trace donor is inferred with high accuracy from the DNA sample. Although even if the current issues are overcome in the future, caution should be taken into account regarding the trait variation within biogeographic groups. Prior incorporation could be considered either in a continental level by averaging all countries that belong to this continent or a population group level. For continents such as Europe, where phenotypic variation is high between populations from different parts, such averaging of the prevalence values might not be beneficial and will likely lead to misspecification of priors. This problem could be solved by increasing the levels of detail in biogeographic ancestry testing through DNA from continental to population or even sub regional level, which is currently not available for the small sets of markers that can be considered in forensic genetics to date.

To sum up, this study provided a first approach on the possible impact of trait prevalence informed priors for a set of specific EVCs. My results showed that such priors have potential to improve the prediction accuracy for EVC prediction, however there is a certain need for accurate prevalence values in order to perform accordingly. Further studies are needed for the collection of such data on a worldwide level and also for the identification of more causal genetic factors. Only in this way we will be able to identify the actual impact that priors have on EVC prediction.

## 12.4 Non-substantial differences in the prediction performance among all classifiers

In this study, I aimed at comparing four ML classifiers for predicting EVCs including eye, hair & skin color. The classifiers that were applied were MLR, SVM, RF and ANN with the set of predictive markers as were previously established for each trait [5-7]. So far, MLR has been widely used for EVC prediction and many models developed are based on it. Motivated by the question whether any of the aforementioned methods outperform the standard MLR, I applied and then compared those classifiers with the standard performance metrics for classification. Comparative analysis was

conducted in two different datasets, the first one including only European samples while the other one comprised of data samples in a worldwide level. The results obtained from this study showed in general that all classifiers tested here performed almost equally well for the three appearance traits. Especially for eye color almost no variations were obtained while for hair and skin color the variations were slightly higher, however they were not substantial. The same pattern of behavior was obtained in both datasets which leads to the conclusion that none of the three classifiers applied here outperformed MLR, at least for the currently known marker sets. This was expected if we consider that European samples represent the majority of the complete dataset and therefore in both cases the model was trained mostly in European samples.

For eye color, I obtained that for brown and blue categories, all four classifiers performed almost identically while for intermediate eye color I observed small deviations for some performance measurements. This is likely explained by the lower prediction performance of intermediate eye color compared with the other two eye color categories, which was also confirmed by previous studies and is probably due to the lack of knowledge on markers that contribute to this pigmentation trait [5-7, 63, 75, 167]. Future studies that aim for the identification of such markers are necessary in order to improve the prediction accuracy of intermediate eye color.

Regarding the prediction of hair color, I obtained slightly higher deviations among the four classification methods, however they were not substantial. Similar behavior was obtained also for skin color prediction, which was the trait with the highest deviations between model performances compared with the previous two. This could be possibly explained by the fact that there is a larger number of predictors contributing for these traits that are either not yet identified or they are some external or environmental factors that so far cannot be incorporated in the prediction models. This can affect the outcomes of the classifiers, meaning that each method might have a different level of sensitivity in detecting the impact of external factors, which subsequently leads to differentiations in the results of the different methods. Another possible explanation is the fact that for hair and skin color, a larger number of predictors are contributing, that affects the ability of the classifiers to pick up local patterns, and in combination with the relatively small datasets this could lead to some random events which subsequently deteriorates the overall model performance.

To give an overview, the systematic comparison of the four classification methods showed non-substantial differences between the model performances. This could be due to the fact that we look at the additive effects of the genetic markers and not at potential interactions. This could be possibly explained by the way that these genetic markers were identified. The identification of such markers is usually done by GWAS which mostly focus on the additive effects of each marker to the trait tested. By looking at possible interaction effects of the predictive markers, some additional information that affects the current prediction performances might be obtained. However, with the currently known markers and due to the fact that ML approaches applied here look implicitly at interaction effects and the already established studies that have shown small effects on SNP-SNP interactions, it is not recommended at least for this stage to incorporate interaction effects. Additional knowledge for predictive markers is required in order to see whether such interactions could actually affect the performance of the different classifiers. Furthermore, according to the

results, it was obtained that the performance metrics of all classifiers reflect the genetic determination of the trait categories. This means that categories such as red hair color which are strongly determined by certain genes tended to have higher performance metrics compared to other categories. Another explanation for the results obtained among the four classifiers, could be the relatively small data sizes used here for EVC prediction. By nature, ML approaches use the information from the data, they 'learn' from it and subsequently use this information in order to classify new observations. That means that the more information is available to the model during training the more accurate the predictions will be. This was also confirmed with the two datasets that I used, where in the complete dataset we had more observations and the performance measurements were overall higher, compared to the smaller European dataset. In ML approaches, large datasets are generally preferable in order to provide more information in the training model and avoid possible random patterns that can often occur in smaller datasets.

To sum up, according to the current findings and the currently known predictive markers, it is suggested that MLR is the most appropriate and preferable method for EVC prediction due to its easier interpretation compared to the other classifiers. MLR gives a clearer insight of the interactions between the features and as well the importance of each feature to the outcomes while for the rest of ML approaches this is difficult to obtain. However, in the future this statement might require re-evaluation, as we make more progress towards the identification of more genetic markers and their possible interactions that contribute to appearance traits.

# 13 Outlook

The results of my PhD showed that there is a proportion of priors that show potential to improve the prediction accuracies of the already existing EVC models. Furthermore, I found that ML approaches including SVM, RF and ANN did not show any substantial differences and performed almost equally well with the conventional method of MLR. However, it is of utmost importance to collect accurate prevalence data on pigmentation traits of different populations in order to see the actual impact of priors on EVC prediction. On the other hand, the identification of novel DNA markers that possibly explain at a higher extent the phenotypic variations of the traits will improve their prediction accuracy.

Based on the preliminary study of the literature review on appearance traits, I managed to collect data on eye color for 16 European and Central Asian countries and on hair color for 7 European countries. For the rest of the appearance traits, data were either not sufficient or non-existent. It is obvious that only with the aforementioned data available, the incorporation of accurate prior values in the EVC prediction model was not feasible at this stage. Therefore, further studies that focus on the spatial distribution of pigmentation traits are necessary in order to see whether the accurate priors do actually improve the current prediction models. In a next step, it would be important to identify more genetic markers that contribute to different appearance traits and have not been identified so far. Since the current approaches for EVC prediction have potential for improvement, the addition of new genetic markers and possibly their interactions might be necessary in order to

achieve that. Large scale GWAS are necessary to be conducted in the future for the identification of such markers. Incorporation of novel genetic markers, their SNP-SNP interactions, epigenetic or other features into the prediction model might not only improve the overall prediction accuracy but also might indicate whether other ML approaches outperform the conventional MLR. Furthermore it would be necessary for future studies to generate and apply larger and well defined datasets in order to obtain more accurate outcomes especially in case of comparing various classifiers. ML approaches are learning from "experience", that means that the more data you provide, the more information is incorporated in the model which subsequently can lead to more accurate predictions. The relatively small data size available for eye, hair and skin color in my study was likely an obstacle in order to see the actual performance of all four classifiers. With the knowledge that is currently available on the genetic markers and the current datasets, the recommended method for EVC prediction is MLR due to the fact is a more interpretable method that provides information on the feature interaction and contribution to the outcome.

Notably, the results obtained throughout my PhD project, despite the limitations of available data that I faced, provide a significant insight of novel approaches and systematic comparison of various established classifiers. These results can motivate further studies on the spatial distribution of appearance traits and the identification of novel genetic markers in order to improve the already existing approaches for EVC prediction.

# 14 Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr. Michael Nothnagel who gave me the opportunity to conduct this PhD project, for his continuous support through it. His immense knowledge helped me significantly throughout these years of research and during the writing of my thesis. Furthermore, I would like to thank him for his motivation and the opportunities he gave me to participate in various conferences that educated me further and made me gain insight into interesting scientific projects. I am very thankful for his help in preparing talks and posters for the conferences and for all the advice he gave me throughout the last three years. I would like to thank him also for reading early drafts of this thesis that due to his experience and excellent writing skills he helped improving it with comments and suggestions. Lastly, I want to thank Michael not only for our excellent collaboration and the knowledge I gained during my PhD, but also for his help in practicing the German language and also his advice and opportunities regarding my professional future.

I would like to thank also Prof. Dr. Peter Schneider and Prof. Dr. Thomas Wiehe for agreeing to be my tutors and for taking time for our meetings and their advice and comments on my PhD thesis.

Furthermore, I would like to thank all members of the VISAGE consortium for our fruitful and smooth collaboration all these three years. Their help was fundamental and it is only with their data that I was able to proceed with my research and complete my PhD. Special thanks to Prof. Dr. Manfred Kayser for his guidance in using the appropriate datasets throughout this study and also for his many suggestions for improving my main investigations. Many thanks also to Dr. Christopher

Phillips for the very good collaboration and exchange of information on the data used regarding the inference of biogeographic ancestry.

Of course, I would like to thank all current and former colleagues namely Dr. Dmitriy Drichel, Dr. George Kanoungi, Oluyomi Adesoji and Tarek Khellaf for their support and help all these years and also for letting me being part of this amazing team. It was amazing working with all of you. I am very grateful for all the warm chocolates we had together and for the reassurance in moments of helplessness according to the moto “life is beautiful”. I want to special thank also Dr. Dmitriy Drichel for accepting to proofread my final thesis for helping me to improve it according to his immense knowledge and experience. Thank you as well for your encouraging words.

I would like also to thank Dr. Gabriele Thorn for her support, encouragement and for standing by my side in difficult moments. Thank you for letting me stay with you the time that I needed. It was lovely. Your loving care and motivation was like a familial support to me. I also have to acknowledge Heinrich Rohde and Khanh Toan Nguyen for providing their technical support when was necessary. Furthermore, I wish to thank all employees of the Cologne Center for Genomics (CCG) that made my working period in the institute and amazing experience.

Last but not least, I would like to thank my family and friends for all the support and encouragement. Thank you for believing in me and that you are always by my side in every step I am making in my life.

# 15 Erklärung

**Erklärung zur Dissertation**
gemäß der Promotionsordnung vom 12. März 2020

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

Teilpublikationen:

**Katsara M.A.**, Nothnagel M. 2019. **True colors: A literature review on the spatial distribution of eye and hair pigmentation**, Forensic Science International Genetics, 39:109-118, published online January 02, 2019. Doi: https://doi.org/10.1016/j.fsigen.2019.01.001

**Katsara M.A.**, Branicki W., Pospiech E., Hysi P., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. **Testing the impact of trait prevalence priors in Bayesian-based genetic prediction modeling of human appearance traits**, Forensic Science International Genetics, 50:102412, published online November 03, 2020. Doi: https://doi.org/10.1016/j.fsigen.2020.102412

**Katsara M.A.**, Branicki W., Walsh S., Kayser M., Nothnagel M., on behalf of the VISAGE Consortium. **Evaluation of supervised machine-learning methods for predicting appearance traits from DNA**, submitted to Forensic Science International Genetics, under revision.

19/01/2021, Maria-Alexandra Katsara,

Datum, Name und Unterschrift

# 16 References


1. Kanchan, T. and K. Krishan, *Personal Identification in Forensic Examinations.* Anthropology, 2013. **2**(1): p. 1-2.
2. Lareu, M., *Short Tandem Repeats*. Second ed2013: Encyclopedia of Forensic Sciences (Second Edition).
3. Marano, L.A. and C. Fridman, *DNA phenotyping: current application in forensic science.* Research and Reports in Forensic Medical Science, 2019. **9**: p. 1-8.
4. Schneider, P., B. Prainsack, and M. Kayser, *The Use of Forensic DNA Phenotyping in Predicting Appearance and Biogeographic Ancestry.* Deutsche Ärzteblatt International, 2019. **116**(51-52): p. 873-880.
5. Walsh, S., et al., *IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information.* Forensic Science International Genetics 2011. **5**(3): p. 170-180.
6. Walsh, S., et al., *The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA.* Forensic Science International Genetics, 2013. **7**(1): p. 98-115.
7. Chaitanya, L., et al., *The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation* Forensic Science International Genetics, 2018. **35**: p. 123-135.
8. Bhavsar, P., et al., *Chapter 12 - Machine Learning in Transportation Data Analytics*2017: Data Analytics for Intelligent Transportation Systems.
9. Rudin, N., et al., *DNA based Identification* Biometrics1996, Boston Springer, Boston, MA.
10. Norrgard, K., *Forensics, DNA Fingerprinting, and CODIS.* Nature Education 2008. **1**(1): p. 35.
11. Giannelli, P., *The DNA Story: An Alternative View.* The Journal of Criminal Law and Criminology, 1997. **88**(1): p. 380-422.
12. Hibbert, M., *DNA databanks: law enforcement's greatest surveillance tool? .* Wake Forest Law Rev., 1999. **34**(3): p. 767-825.
13. Angier, N., *SCIENTIST AT WORK: Mary-Claire King; Quest for Genes and Lost Children.* The New York Times, 1993: p. 1.
14. Erlich, Y., et al., *Identity inference of genomic data using long-range familial searches.* Science, 2018. **362**(6415): p. 690-694.
15. Lynch, M., *God's signature: DNA profiling, the new gold standard in forensic science.* Endeavour, 2003. **27**(2): p. 93-97.
16. Wyner, N., M. Barash, and D. McNevin, *Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype.* Frontiers in Genetics, 2020.
17. *FBI CODIS Core STR loci* National Institute of Standards and Technology 2015; Available from: https://strbase.nist.gov/fbicore.htm.
18. Butler, J.M., *Genetics and genomics of core STR loci used in human identity testing.* J Forensic Sci., 2006. **51**(2): p. 253-265.
19. Mulero, J.J., et al., *Development and validation of the AmpFlSTR MiniFiler PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA* J Forensic Sci., 2008. **53**(4): p. 838-852.
20. Greenspoon, S.A., et al., *Validation and implementation of the PowerPlex 16 BIO System STR multiplex for forensic casework* J Forensic Sci., 2004. **49**(1): p. 71-80.
21. Thakur, M., et al., *Functional validation of human-specific PowerPlex® 21 System (Promega, USA) in chimpanzee (Pan troglodytes).* BMC Research Notes, 2018. **695**.

22. Chakraborty, R., et al., *The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems.* Electophoresis, 1999. **20**(8): p. 1682-1696.
23. Urquhart, A., et al., *Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers.* Int J Legal Med, 1994. **107**(1): p. 13-20.
24. Education, N., *"single-nucleotide polymorphism / SNP | Learn Science at Scitable".* www.nature.com 2015.
25. Yang, Y., B. Xie, and J. Yan, *Application of Next-generation Sequencing Technology in Forensic Science.* Genomics, Proteomics & Bioinformatics, 2014. **12**(5): p. 190-197.
26. Samuel, G. and B. Prainsack, *Societal, ethical, and regulatory dimensions of forensic DNA phenotyping. VISAGE.* 2019.
27. Toom, V., et al., *Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively: Reply to 'Forensic DNA phenotyping: Predicting human appearance from crime scene material for investigative purposes' by Manfred Kayser* Forensic Science International Genetics, 2016. **22**: p. e1-e4.
28. Dror, I.E. and G. Hampikian, *Subjectivity and bias in forensic DNA mixture interpretation.* Science & Justice, 2011. **51**(4): p. 204-208.
29. Wallace, H.M., et al., *Forensic DNA databases–Ethical and legal standards: A global review.* Egyptian Journal of Forensic Sciences, 2014. **4**(3): p. 57-63.
30. Cho, M.K. and P. Sankar, *Forensic genetics and ethical, legal and social implications beyond the clinic.* Nature Genetics, 2004. **S8-S12**.
31. Samuel, G. and B. Prainsack, *Forensic DNA phenotyping in Europe: views "on the ground" from those who have a professional stake in the technology.* New Genetics and Society, 2018. **38**(2): p. 119-141.
32. Koops, B.-J. and M. Schellekens, *Forensic DNA phenotyping: Regulatory issues.* Science and Technology Law Review, 2008: p. 158-202.
33. Weichert, T., *Genetische Forensik und Datenschutz.* Zeitschrift für Bürgerrechte und Gesellschaftspolitik. Journal of Civil Rights and Social Policy, 2017. **218**(2): p. 123-134.
34. Scudder, N., et al., *Massively parallel sequencing and the emergence of forensic genomics: Defining the policy and legal issues for law enforcement.* Sci Justice, 2018b. **58**(2): p. 153-158.
35. Turna, R., *Push for Forensic DNA Phenotyping, Ancestry Testing in Germany Raises Discrimination Concerns* Geome Web, 2018.
36. Kayser, M. and P. Schneider, *DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations.* Forensic Science International Genetics, 2009. **3**(3): p. 154-161.
37. de Cerqueira, C.C.S., et al., *Predicting Physical Features and Diseases by DNA Analysis: Current Advances and Future Challenges.* Journal of Forensic Research, 2016.
38. Kayser, M., *Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes* Forensic Science International Genetics, 2015. **18**: p. 33-48.
39. MacLean, C.E., *Creating a Wanted Poster from a Drop of Blood: Using DNA Phenotyping to Generate an Artist's Rendering of an Offender Based Only on DNA Shed at the Crime Scene.* Hamline Law Review, 2014. **36**(3): p. 357-386.
40. Smith, M. and G.F. Urbas, *Regulating new forms of forensic DNA profiling under Australian legislation: familial matching and DNA phenotyping.* Australian Journal of Forensic Sciences, 2012. **44**(1): p. 63-81.

41. Ossorio, P.N., *About face: forensic genetic testing for race and visible traits* Journal of law, medicine & ethics, 2006. **34**(2): p. 277-292.
42. Scudder, N., et al., *Forensic DNA phenotyping: Developing a model privacy impact assessment* Forensic Science International Genetics, 2018. **34**: p. 222-230.
43. Barreiro, L.B., et al., *Natural selection has driven population differentiation in modern humans.* Nature Genetics, 2008. **40**: p. 340-345.
44. Norrgard, K. and J. Schultz, *Using SNP Data to Examine Human Phenotypic Differences.* Nature Education, 2008. **1**(1): p. 85.
45. Guo, J., et al., *Variation and signatures of selection on the human face.* Journal of Human Evolution, 2014. **75**: p. 143-152.
46. Sobrino, B. and A. Carracedo, *SNP Typing in Forensic Genetics*. Forensic DNA Typing Protocols Methods in Molecular Biology. Vol. 297. 2005: Humana Press.
47. Shastry, B.S., *SNPs: impact on gene function and phenotype.* Methods in Molecular Biology, 2009. **578**: p. 3-22.
48. Nachman, M.W., *Single nucleotide polymorphisms and recombination rate in humans.* Trends in Genetics, 2001. **17**(9): p. 481-485.
49. Turakulov, R. and S. Easteal, *Number of SNPS loci needed to detect population structure* Human Heredity, 2003. **55**(1): p. 37-45.
50. Sampson, J., et al., *Selecting SNPs to Identify Ancestry.* Ann Hum Genet, 2011. **75**(4): p. 539-553.
51. Butler, J.M., *Fundamentals of forensic DNA typing*2010: Elsevier/Academic Press.
52. Liu, F., et al., *Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair.* Hum Mol Genet, 2018. **27**(3): p. 559-575.
53. Sulem, P., et al., *Genetic determinants of hair, eye and skin pigmentation in Europeans* Nature Genetics, 2007. **39**(12): p. 1443-1452.
54. Pirastu, N., et al., *GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk.* Nature Communications 2017. **8**.
55. Yap, C., et al., *Dissection of genetic variation and evidence for pleiotropy in male pattern baldness.* Nature Communications, 2018. **9**.
56. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry* Hum Mol Genet, 2018. **27**(20): p. 3641-3649.
57. Chan, Y., et al., *Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development.* Am J Hum Genet, 2015. **96**(5): p. 695-708.
58. Beals, R.L. and H. Hoijer, *An introduction to anthropology*. 3rd edition ed1965: New York : Macmillan, ©1965.
59. Kayser, M., et al., *Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene* Am J Hum Genet, 2008. **82**(2): p. 411-423.
60. Eiberg, H., et al., *Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression* Human Genetics, 2008. **123**(2): p. 177-187.
61. Frost, P., *European hair and eye color: A case of frequency-dependent sexual selection?* Evolution and Human Behavior, 2006. **27**(2): p. 85-103.
62. Cavalli-Sforza, L.L., P. Menozzi, and A. Piazza, *The History and Geography of Human Gene.* The Journal of Asian Studies, 1995.
63. Liu, F., et al., *Eye color and the prediction of complex phenotypes from genotypes.* Current Biology, 2009. **19**(5).

64. Sturm, R.A., et al., *A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color* Am J Hum Genet, 2008. **82**(2): p. 424-431.
65. Han, J., et al., *A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation.* PLoS Genetics, 2008. **4**(5).
66. Sulem, P., et al., *Two newly identified genetic determinants of pigmentation in Europeans* Nature Genetics, 2008. **40**(7): p. 835-837.
67. Kanetsky, P.A., et al., *A polymorphism in the agouti signaling protein gene is associated with human pigmentation.* Am J Hum Genet, 2002. **70**(3): p. 770-775.
68. Duffy, D.L., et al., *A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of OCA2 Explains Most Human Eye-Color Variation.* Am J Hum Genet, 2007. **80**(2): p. 241-252.
69. Zhu, G., et al., *A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q* Twin research : the official journal of the International Society for Twin Studies, 2004. **7**(2): p. 197-210.
70. Posthuma, D., et al., *Replicated Linkage for Eye Color on 15q Using Comparative Ratings of Sibling Pairs.* Behavior Genetics, 200.
71. Frudakis, T., et al., *Sequences associated with human iris pigmentation.* Genetics, 2003. **165**(4): p. 2071-2083.
72. Rebbeck, T.R., et al., *P gene as an inherited biomarker of human eye color.* Cancer Epidemiol Biomarkers Prev, 2002. **11**(8): p. 782-784.
73. White, D. and M. Rabago-Smith, *Genotype–phenotype associations and human eye color.* Journal of Human Genetics 2011. **56**: p. 5-7.
74. Liu, F., et al., *Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci.* PLoS Genetics, 2010.
75. Walsh, S., et al., *DNA-based eye colour prediction across Europe with the IrisPlex system* Forensic Science International Genetics, 2012. **6**: p. 330-340.
76. Pospiech, E., et al., *Gene-gene interactions contribute to eye colour variation in humans* Journal of Human Genetics, 2011. **56**(6): p. 447-455.
77. Pietroni, C., et al., *The effect of gender on eye colour variation in European populations and an evaluation of the IrisPlex prediction model.* Forensic Science International Genetics, 2014. **11**: p. 1-6.
78. Martinez-Cadenas, C., et al., *Gender is a major factor explaining discrepancies in eye colour prediction based on HERC2/OCA2 genotype and the IrisPlex model* Forensic Science International Genetics, 2013. **7**(4): p. 453-460.
79. Siewierska-Gorska, A., et al., *Association of five SNPs with human hair colour in the Polish population.* HOMO - Journal of Comparative Human Biology 2016.
80. Rees, J., *Genetics of hair and skin color* Annual Review of Genetics, 2003. **37**: p. 67-90.
81. Shekar, S.N., et al., *Linkage and association analysis of spectrophotometrically quantified hair color in Australian adolescents: the effect of OCA2 and HERC2* The Journal of investigative dermatology, 2008. **128**(12): p. 2807-2814.
82. Mengel-From, J., et al., *Genetic determinants of hair and eye colours in the Scottish and Danish populations.* BMC Genetics, 2009.
83. Valverde, P., et al., *Variants of the melanocyte–stimulating hormone receptor gene are associated with red hair and fair skin in humans.* Nature Genetics, 1995. **11**: p. 328-330.
84. Branicki, W., et al., *Association of the SLC45A2 gene with physiological human hair colour variation.* Journal of Human Genetics, 2008. **53**: p. 966-971.
85. Graf, J., R. Hodgson, and A. van Daal, *Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation* Hum Mutat, 2005. **25**(3): p. 278-284.

86. Grimes, E.A., et al., *Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype.* Forensic Sci Int, 2001. **122**(2-3): p. 124-129.
87. Branicki, W., et al., *Model-based prediction of human hair color using DNA variants.* Human Genetics, 2011. **129**(4): p. 443-454.
88. Kenny, E.E., et al., *Melanesian blond hair is caused by an amino acid change in TYRP1* Science (New York, N.Y.), 2012. **336**(6081): p. 554.
89. Kukla-Bartoszek, M., et al., *Investigating the impact of age-depended hair colour darkening during childhood on DNA-based hair colour prediction with the HIrisPlex system.* Forensic Science International Genetics, 2018. **36**: p. 26-33.
90. Deng, L. and S. Xu, *Adaptation of human skin color in various populations.* Hereditas, 2018. **155**: p. 1.
91. Costin, G.-E. and V.J. Hearing, *Human skin pigmentation: melanocytes modulate skin color in response to stress.* The FASEB Journal, 2007. **21**(4): p. 976-994.
92. Brenner, M. and V.J. Hearing, *The Protective Role of Melanin Against UV Damage in Human Skin.* Photochem Photobiol, 2008. **84**(3): p. 539-549.
93. Stokowski, R.P., et al., *A Genomewide Association Study of Skin Pigmentation in a South Asian Population.* Am J Hum Genet, 2007. **81**(6): p. 1119-1132.
94. Jacobs, L.C., et al., *Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans* Human Genetics, 2013. **132**(2): p. 147-158.
95. Zaorska, K., P. Zawierucha, and M. Nowicki, *Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches.* Human Genetics, 2019. **138**: p. 635-647.
96. Beleza, S., et al., *Genetic Architecture of Skin and Eye Color in an African-European Admixed Population.* PLoS Genetics, 2013.
97. Wolf Horrell, E.M., M.C. Boulanger, and J.A. D'Orazio, *Melanocortin 1 Receptor: Structure, Function, and Regulation.* Frontiers in Genetics, 2016. **7**: p. 95.
98. Miller, C.T., et al., *Cis-regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans.* HHS Author Manuscripts, 2007. **131**(6): p. 1179-1189.
99. Loussouarn, G., et al., *Diversity in human hair growth, diameter, colour and shape. An in vivo study on young adults from 24 different ethnic groups observed in the five continents* European Journal of Dermatology, 2016. **26**(2): p. 144-154.
100. Medland, S.E., G. Zhu, and N.G. Martin, *Estimating the heritability of hair curliness in twins of European ancestry* Twin research and human gentics: the official journal of the International Society for Twin Studies, 2009. **12**(5): p. 514-518.
101. Loussouarn, G., et al., *Worldwide diversity of hair curliness: a new method of assessment.* International Journal of Dermatology, 2007.
102. Pospiech, E., et al., *Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA.* Forensic Science International Genetics, 2018. **37**: p. 241-251.
103. Eriksson, N., et al., *Web-based, participant-driven studies yield novel genetic associations for common traits* PLoS Genetics, 2010. **6**(6).
104. Medland, S.E., et al., *Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans.* Am J Hum Genet, 2009. **85**(5): p. 750-755.
105. Adhikari, K., et al., *A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features.* Nature Communications, 2016. **7**.

106. Wu, S., et al., *Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations* Human Genetics, 2016. **135**(11): p. 1279-1286.
107. Fujimoto, A., et al., *A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness* Hum Mol Genet, 2008. **17**(6): p. 835-843.
108. Stecksen-Blicks, C., et al., *Hair shaft structures in EDAR induced ectodermal dysplasia.* BMC Medical Genetics, 2015. **79**.
109. Pospiech, E., et al., *Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans.* Forensic Science International: Genetics, 2015. **19**: p. 280-288.
110. Praetorius, C., R.A. Sturm, and E. Steingrimsson, *Sun-induced freckling: ephelides and solar lentigines.* Pigment Cell Melanoma Research, 2014.
111. Plensdorf, S. and J. Martinez, *Common pigmentation disorders* Am Fam Physician, 2009. **79**(2): p. 109-116.
112. Bishop, D.T., F. Demenais, and J.A. Newtin Bishop, *Genome-wide association study identifies three loci associated with melanoma risk.* Nature Genetics, 2009. **41**: p. 920-925.
113. Gudbjartsson, D.F., P. Sulem, and K. Stefansson, *ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma.* Nature Genetics, 2008. **40**: p. 886-891.
114. Kosiniak-Kamysz, A., et al., *Increased risk of developing cutaneous malignant melanoma is associated with variation in pigmentation genes and VDR, and may involve epistatic effects.* Melanoma Research 2014. **24**(4): p. 388-396.
115. Kosiniak-Kamysz, A., et al., *Potential association of single nucleotide polymorphisms in pigmentation genes with the development of basal cell carcinoma.* The Journal of Dermatology, 2012.
116. Rees, J., *The genetics of sun sensitivity in humans.* Am J Hum Genet, 2004. **75**(5): p. 739-751.
117. Bastiaens, M., et al., *The melanocortin-1-receptor gene is the major freckle gene* Hum Mol Genet, 2001. **10**(16): p. 1707-1708.
118. Garcia-Borron, J.C., B.L. Sanchez-Laorden, and C. Jimenez-Cervantes, *Melanocortin-1 receptor structure and functional regulation* Pigment Cell Melanoma Research, 2005. **18**(6): p. 393-410.
119. Hernando, B., et al., *Genetic determinants of freckle occurrence in the Spanish population: Towards ephelides prediction from human DNA samples.* Forensic Science International Genetics, 2018. **33**: p. 38-47.
120. Kukla-Bartoszek, M., et al., *DNA-based predictive models for the presence of freckles.* Forensic Science International Genetics, 2019. **42**: p. 252-259.
121. Nyholt, D.R., et al., *Genetic basis of male pattern baldness.* J Invest Dermatol, 2003. **121**(6): p. 1561-1564.
122. Cash, T.F., *The psychosocial consequences of androgenetic alopecia: a review of the research literature* The British Journal of dermatology, 1999. **141**(3): p. 398-405.
123. Christoffersen, M., et al., *Visible age-related signs and risk of ischemic heart disease in the general population: a prospective cohort study.* Circulation, 2014. **129**(9): p. 990-998.
124. Shahar, E., et al., *Baldness and myocardial infarction in men: the atherosclerosis risk in communities study* Am J Epidemiology, 2008. **167**(6): p. 676-683.
125. Trieu, N. and G.D. Eslick, *Alopecia and its association with coronary heart disease and cardiovascular risk factors: a meta-analysis.* International Journal of Cardiology, 2014. **176**(3): p. 687-695.
126. Cremers, R.G., et al., *Androgenic alopecia is not useful as an indicator of men at high risk of prostate cancer.* European Journal of Cancer, 2016. **46**(18): p. 3294-3299.

127. Zhou, C.K., et al., *Male Pattern Baldness in Relation to Prostate Cancer-Specific Mortality: A Prospective Analysis in the NHANES I Epidemiologic Follow-up Study.* Am J Epidemiology, 2016. **183**(3): p. 210-217.
128. Zhou, C.K., et al., *Relationship between male pattern baldness and the risk of aggressive prostate cancer: an analysis of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial* J Clin Oncol., 2015. **33**(5): p. 419-425.
129. Rexbye, H., et al., *Hair loss among elderly men: etiology and impact on perceived age* The Journals of Gerontology. Series A, Biological sciences and medical sciences, 2005. **60**(8): p. 1077-1082.
130. Brockschmidt, F.F., et al., *Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness* The British Journal of dermatology, 2011. **165**(6): p. 1293-1302.
131. Brockschmidt, F.F., et al., *Fine mapping of the human AR/EDA2R locus in androgenetic alopecia* The British Journal of dermatology, 2010. **162**(4): p. 899-903.
132. Marcinska, M., et al., *Evaluation of DNA Variants Associated with Androgenetic Alopecia and Their Potential to Predict Male Pattern Baldness.* PLOS One 2015.
133. Gupta, M. and V. Mysore, *Classifications of Patterned Hair Loss: A Review.* J Cutan Aesthet Surg, 2016. **9**(1): p. 3-12.
134. Liu, F., et al., *Common DNA variants predict tall stature in Europeans.* Human Genetics, 2014. **133**: p. 587-597.
135. Marouli, E., et al., *Rare and low-frequency coding variants alter human adult height.* Nature, 2017. **542**(7640).
136. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nature Genetics, 2014. **46**(11): p. 1173-1186.
137. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature 2010. **467**(7317): p. 832-838.
138. Macgregor, S., et al., *Bias, precision and heritability of self-reported and clinically measured height in Australian twins* Human Genetics, 2006. **120**(4): p. 571-580.
139. Carmichael, C.M. and M. McGue, *A cross-sectional examination of height, weight, and body mass index in adult twins* The Journals of Gerontology. Series A, Biological sciences and medical sciences, 1995. **50**(4): p. B237-244.
140. Phillips, K. and A.P.J. Matheny, *Quantitative genetic analysis of longitudinal trends in height: preliminary results from the Louisville Twin Study* Acta Genet Med Gemellol (Roma), 1990. **39**(2): p. 143-163.
141. Liu, F., et al., *Update on the predictability of tall stature from DNA markers in Europeans.* Forensic Science International Genetics, 2019. **42**: p. 8-13.
142. Russell, S. and P. Norvig, *Artificial Intelligence: A Modern Approach*2009: Prentice Hall PressOne Lake Street Upper Saddle River, NJUnited States.
143. Hinton, G. and T.J. Sejnowski, *Unsupervised Learning: Foundations of Neural Computation*1999: The MIT Press.
144. Bunge, J.A. and D.H. Judson, *Data Mining.* Encyclopedia of Social Measurement, 2005: p. 617-624.
145. Mitchell, T.M., *Machine Learning*1997, West 19th Street New York, NYUnited States: McGraw-Hill, Inc.Professional Book Group 11
146. Boser, B., I. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers.*1992, Pittsburgh Proceedings of the Fifth Annual Workshop on Computational Learning Theory.
147. Vapnik, V., *The Support Vector method. .* Artificial Neural Networks — ICANN'97. ICANN 1997, 1997. **1327**.

148. Breiman, L., *Random Forests* Machine Learning, 2001. **45**: p. 5-32.
149. Dev, V.A. and M.R. Eden, *Gradient Boosted Decision Trees for Lithology Classification.* Computer Aided Chemical Engineering, 2019. **47**: p. 113-118.
150. Probst, P. and A.-L. Boulesteix, *To Tune or Not to Tune theNumber of Trees in Random Forest.* Journal of Machine Learning Research, 2018. **18**: p. 1-18.
151. Breiman, L., *Bias, variance and arcing classifiers.* 1996.
152. Liaw, A. and M. Wiener, *Classification and Regression by randomForest.* Computer Science, 2007.
153. Oshiro, T.M., P. Santoro Perez, and J.A. Baranauskas, *How Many Trees in a Random Forest?* Machine Learning and Data Mining in Pattern Recognition. Vol. 7376. 2012, Berlin, Heidelberg: Springer, Berlin, Heidelberg.
154. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*2009: Springer, New York, NY.
155. Lek, S. and Y.S. Park, *Artificial Neural Networks.* Encyclopedia of Ecology, 2008: p. 237-245.
156. Sengupta, N., M. Sahidullah, and G. Saha, *Lung sound classification using cepstral-based statistical features.* Computers in Biology and Medicine, 2016. **75**: p. 118-129.
157. Balabin, R.M. and E.I. Lomakina, *Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies.* The Journal of Chemical Physics, 2009. **131**.
158. French, J., *The time traveller's CAPM.* Investment Analysts Journal, 2017. **46**(2): p. 81-96.
159. Joyce, J., *Bayes' Theorem*. Spring 2019 ed. The Stanford Encyclopedia of Philosophy2019: Metaphysics Research Lab, Stanford University.
160. McLachlan, G., *Discriminant Analysis and Statistical Pattern Recognition* 2004, New Jersey: John Wiley & Sons, Inc.
161. Nothnagel, M., *Klassifikationsverfahren der Diskriminanzanalyse Eine vergleichende und integrierende Übersicht*, in *Mathematisch-Naturwissenschaftliche Fakultät II*1999, Hamboldt-Universität zu Berlin: Berlin.
162. Duda, R., P. Hart, and D. Stork, *Pattern Classification*2001, New York: Wiley.
163. Ruiz, Y., et al., *Further development of forensic eye color predictive tests.* Forensic Science International Genetics, 2013. **7**: p. 28-40.
164. Maroñas, O., et al., *Development of a forensic skin colour predictive test* Forensic Science International Genetics, 2014. **13**: p. 34-44.
165. Söchtig, J., et al., *Exploration of SNP variants affecting hair colour prediction in Europeans* International Journal of Legal Medicine, 2015. **129**(5): p. 963-975.
166. Hart, K.L., et al., *Improved eye- and skin-color prediction based on 8 SNPs.* Croat Med J, 2013. **54**(3): p. 248-256.
167. Allwood, J.S. and S. Harbison, *SNP model development for the prediction of eye colour in New Zealand.* Forensic Science International Genetics, 2013. **7**(4): p. 444-452.
168. Virchow, R., *Gesammtbericht über die von der Deutschen Anthropologischen Gesellschaft veranlassten Erhebungen über die Farbe der Haut, der Haare und der Augen der Schulkinder in Deutschland.* **CCXII**: p. 232.
169. Hysi, P.G., et al., *Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability* Nature Genetics, 2018. **50**(5): p. 652-656.
170. Visconti, A., et al., *Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure.* Nature Communications, 2018. **9**.